

# UCSF

## UC San Francisco Previously Published Works

### Title

Retrospective comparative effectiveness research: Will changing the analytical methods change the results?

### Permalink

<https://escholarship.org/uc/item/4vv097kb>

### Journal

International journal of cancer, 150(12)

### ISSN

0020-7136

### Authors

Zaorsky, Nicholas G  
Wang, Xi  
Lehrer, Eric J  
[et al.](#)

### Publication Date

2022-06-01

### DOI


10.1002/ijc.33946

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

# Retrospective comparative effectiveness research: Will changing the analytical methods change the results?

Nicholas G. Zaorsky<sup>1,2</sup>  | Xi Wang<sup>2</sup> | Eric J. Lehrer<sup>3</sup> | Leila T. Tchelebi<sup>4,5</sup> | Andrew Yeich<sup>1,2</sup> | Vinay K. Prasad<sup>6</sup> | Vernon M. Chinchilli<sup>2</sup> | Ming Wang<sup>2</sup>

<sup>1</sup>Department of Radiation Oncology, University Hospitals Seidman Cancer Center, Case Western Reserve School of Medicine, Cleveland, Ohio, USA

<sup>2</sup>Department of Public Health Sciences, Penn State College of Medicine, Hershey, Pennsylvania, USA

<sup>3</sup>Department of Radiation Oncology, Icahn School of Medicine at Mount Sinai, New York, New York, USA

<sup>4</sup>Department of Radiation Medicine, Zucker School of Medicine, Hempstead, New York, USA

<sup>5</sup>Department of Radiation Medicine, Northwell Health Cancer Institute, Mount Kisco, New York, USA

<sup>6</sup>Department of Medical Oncology, UCSF, San Francisco, California, USA

## Correspondence

Nicholas G. Zaorsky, Department of Radiation Oncology, Penn State Cancer Institute, 500 University Drive, Hershey, PA 17033, USA.

Email: nzaorsky@pennstatehealth.psu.edu; nicholaszaorsky@gmail.com

## Funding information

NGZ is supported by startup funding from Penn State Cancer Institute and Penn State College of Medicine. NGZ is supported by the National Institutes of Health Grant LRP 1L30 CA231572-01. NGZ is supported by the American Cancer Society—Tri State CEOs Against Cancer Clinician Scientist Development Grant, CSDG-20-013-01-CCE. NGZ received remuneration from Springer Nature for his textbook, Absolute Clinical Radiation Oncology Review.

## Abstract

In medicine, retrospective cohort studies are used to compare treatments to one another. We hypothesize that the outcomes of retrospective comparative effectiveness research studies can be heavily influenced by biostatistical analytic choices, thereby leading to inconsistent conclusions. We selected a clinical scenario currently under investigation: survival in metastatic prostate, breast or lung cancer after systemic vs systemic + definitive local therapy. We ran >300 000 regression models (each representing a publishable study). Each model had various forms of analytic choices (to account for bias): propensity score matching, left truncation adjustment, landmark analysis and covariate combinations. There were 72 549 lung, 14 904 prostate and 13 857 breast cancer patients included. In the most basic analysis, which omitted propensity score matching, left truncation adjustment and landmark analysis, all of the HRs were <1 (generally, 0.60-0.95, favoring addition of local therapy), with all *P-values* <.001. Left truncation adjustment landmark analysis produced results with nonsignificant *P-values*. The combination of propensity score matching, left truncation adjustment, landmark analysis and covariate combinations generally produced *P-values* that were >.05 and/or HRs that were >1 (favoring systemic therapy alone). The use of more statistical methods to reduce the selection bias caused reported HR ranges to approach 1.0. By varying analytic choices in comparative effectiveness research, we generated contrary outcomes. Our results suggest that some retrospective observational studies may find a treatment improves outcomes for patients, while another similar study may find it does not, simply based on analytical choices.

## KEYWORDS

cohort studies, comparative effectiveness research, propensity score, retrospective, selection bias

## What's new?

While randomized controlled trials and retrospective observational studies are standard approaches in comparative effectiveness research, discordance may exist when the two approaches are applied to the same study question. Here, the authors varied analytical choices

for three clinical effectiveness research scenarios to evaluate impacts on conclusions regarding survival in metastatic prostate, breast or lung cancer following local therapy. Analyses show that variations in analytical approach significantly affect hazard ratio, with potential for hundred log-fold differences in *P* values. The tendency toward bias in comparative effectiveness research may be mitigated through requirements for prespecified analytic plans and protocols in journal publications.

## 1 | INTRODUCTION

Comparative effectiveness research evaluates the efficacy of one treatment relative to another, treatment A vs treatment B. In medicine, the gold standard for comparative effectiveness research is a randomized controlled trial.<sup>1-3</sup> However, randomized controlled trials are not always possible, and investigators often use retrospective observational cohort studies to answer comparative effectiveness questions.

There is disagreement between the results of randomized controlled trials and observational studies on the same question. For example, Soni et al<sup>4</sup> found no significant correlation between the hazard ratio (HR) estimates reported by observational studies and randomized trials (concordance correlation coefficient, 0.083; 95% confidence interval [CI], 0.068-0.230). In another study, Kumar et al<sup>5</sup> collected 141 randomized controlled trials in national cancer treatment guidelines and performed their own observational studies, with patient cohorts that match randomized controlled trial study populations. Propensity-weighted HRs for overall survival fell outside the 95% CIs of their randomized trial counterparts 36% of the time (with 64% falling within).

The purpose of the present work was to evaluate whether conclusions reached by comparative effectiveness research with respect to efficacy of a given treatment could be modified by varying the biostatistical methods employed, such as inclusion and exclusion criteria of study population, combinations of covariates in adjusted analyses, selection bias in the study population and immortal time bias.<sup>6,7</sup> We hypothesized that the conclusions of retrospective comparative effectiveness research were sensitive to these analytic choices, based on prior work in other fields.<sup>6,8</sup>

To test our hypothesis, we selected three clinical comparative effectiveness research scenarios in oncology, all related to the idea of local control (ie, treatment of the original primary tumor with surgery and/or radiotherapy) in the setting of metastatic cancer of prostate, female breast and lung. These clinical scenarios were selected because the backbone of therapy for metastatic cancer is systemic therapy (eg, chemotherapy, hormones), and treatment of the initial primary tumor has generally not been recommended.<sup>9</sup> While several prospective studies have generally failed to show a benefit to the addition of local therapy,<sup>10,11</sup> retrospective studies have sometimes shown dramatic improvements in survival when local therapy is added to systemic treatments for patients with metastatic cancer.<sup>12,13</sup>

## 2 | METHODS

We used the National Cancer Database (NCDB), a hospital-based cancer registry that collects data from American College of Surgeons-Commission on Cancer accredited facilities. The database is sponsored by the American College of Surgeons and the American Cancer Society and is recognized as the largest cancer registry worldwide, with over 34 million patients. It includes 70% of all malignant cancers diagnosed in the United States.<sup>14</sup> The NCDB records patient demographics, comorbidities, tumor characteristics and overall survival and contains information regarding therapies delivered during the first course of treatment, including surgery, radiation therapy, immunotherapy and chemotherapy.<sup>15</sup>

We included all patients with metastatic breast, prostate or lung cancer. In the systemic therapy group of each disease site, patients received hormone therapy (for prostate cancer), endocrine therapy (for breast cancer) or chemotherapy (for any of the three of the cancers). In the prostate cancer local therapy group, patients additionally received a prostatectomy or high dose radiotherapy to the prostate gland. In the female breast cancer local therapy group, women additionally received mastectomy. In the lung cancer local therapy group, patients additionally received high dose radiotherapy or lobectomy. Details of selection criteria are shown in Supporting Information File S1.

To modify biostatistical methods, we compared methods that can be divided into three categories, including (a) without propensity score matching vs with propensity score matching; (b) with and without propensity score matching, plus or minus incorporation of left-truncation only; (c) with and without propensity score matching, plus or minus incorporating landmark time points (1-, 6-, 12- and 24-month). To adjust for selection bias<sup>16-18</sup> in selecting study population into the systemic and local therapy group and systemic therapy only group, we apply propensity score matching approaches to mimic the randomized clinical trials.<sup>16-18</sup> A detailed summary of covariates is provided in Supporting Information File S2. We estimate the propensity score, that is, the probability of receiving the systemic and local therapy, with a logistic regression model adjusting for all available covariates related to treatment assignment and the survival outcome. We include all the covariates for the Cox regression.

In the analysis of treatment (ie, systemic + local vs systemic therapy) effect on all-cause survival, we utilized two approaches to address this immortal time bias (illustrated in Supporting Information File S3).<sup>7,19-21</sup> In the first approach, we define left truncation time as time to treatment that is time from diagnosis to systemic therapy or local therapy, whichever happened first and apply the counting process style input in R.<sup>20</sup> In the second approach, we use the “landmark method” to adjust for immortal time bias.<sup>21</sup> Given some prespecified

time point (ie, 1-, 6-, 12- and 24-month), labeled as “landmark,” this method includes patients with time from diagnosis to treatment within the landmark and survived that time point into data analysis. With this method, the conditional survival of the two treatment arms is comparable. Finally, since different covariate combinations may be used in retrospective studies, we quantify the variability of results obtained from these models, called “vibration of effects” using volcano plots,<sup>6</sup> with the  $-\log_{10}(P\text{-value})$  plotted vs the obtained HR.

### 3 | RESULTS

There were 101 310 patients included in the analysis. Among these, 72 549 had metastatic lung cancer, 14 904 had metastatic prostate cancer and 13 857 had metastatic breast cancer. In the lung cancer group, 61 831 patients received systemic therapy and 10 718 received systemic + local therapy; for prostate cancer, 11 737 received systemic while 3167 received systemic + local therapy; for

**TABLE 1** Characteristics of patients with metastatic disease treated with systemic vs systemic and local therapy

	Prostate		Lung		Breast	
	Systemic	Systemic and local	Systemic	Systemic and local	Systemic	Systemic and local
n	11 737	3167	61 831	10 718	9968	3889
Race (%)						
Black	2251 (19.2)	510 (16.1)	7360 (11.9)	1310 (12.2)	2049 (20.6)	721 (18.5)
White	9103 (77.6)	2543 (80.3)	51 894 (83.9)	9069 (84.6)	7486 (75.1)	3011 (77.4)
Other	383 (3.3)	114 (3.6)	2577 (4.2)	339 (3.2)	433 (4.3)	157 (4.0)
T stage (%)						
0/X	3676 (31.3)	457 (14.4)	10 040 (16.2)	852 (7.9)	1598 (16.0)	324 (8.3)
1	2823 (24.1)	966 (30.5)	8438 (13.6)	1136 (10.6)	1167 (11.7)	508 (13.1)
2	2661 (22.7)	632 (20.0)	15 933 (25.8)	2648 (24.7)	2327 (23.3)	1207 (31.0)
3	1233 (10.5)	439 (13.9)	10 558 (17.1)	2449 (22.8)	1399 (14.0)	643 (16.5)
4	1344 (11.5)	673 (21.3)	16 862 (27.3)	3633 (33.9)	3477 (34.9)	1207 (31.0)
Sex (%)						
Male	11 737 (100.0)	3167 (100.0)	33 030 (53.4)	6019 (56.2)	0 (0)	0 (0)
Female	0 (0)	0 (0)	28 801 (46.6)	4699 (43.8)	9968 (100.0)	3889 (100.0)
Insurance status (%)						
No ins.	675 (5.8)	170 (5.4)	2550 (4.1)	504 (4.7)	749 (7.5)	182 (4.7)
Private ins./managed care	2963 (25.2)	863 (27.2)	19 904 (32.2)	3765 (35.1)	4467 (44.8)	1932 (49.7)
Medicaid	833 (7.1)	209 (6.6)	4722 (7.6)	981 (9.2)	1601 (16.1)	592 (15.2)
Medicare	7158 (61.0)	1888 (59.6)	33 932 (54.9)	5292 (49.4)	3068 (30.8)	1148 (29.5)
Other gov. ins.	108 (0.9)	37 (1.2)	723 (1.2)	176 (1.6)	83 (0.8)	35 (0.9)
Adults in patient's zip code without high school diploma (%)						
21% or more	2209 (18.8)	600 (18.9)	10 447 (16.9)	1975 (18.4)	1936 (19.4)	772 (19.9)
13.0%-0.9%	3020 (25.7)	839 (26.5)	16 779 (27.1)	3123 (29.1)	2686 (26.9)	1070 (27.5)
7.0%-12.9%	3771 (32.1)	1001 (31.6)	20 974 (33.9)	3493 (32.6)	3133 (31.4)	1224 (31.5)
Less than 7.0%	2737 (23.3)	727 (23.0)	13 631 (22.0)	2127 (19.8)	2213 (22.2)	823 (21.2)
Average household income in patient's zip code (%)						
Less than \$38 000	2340 (19.9)	617 (19.5)	11 563 (18.7)	2266 (21.1)	2013 (20.2)	805 (20.7)
\$38 000-\$47 999	2844 (24.2)	723 (22.8)	15 206 (24.6)	2791 (26.0)	2323 (23.3)	928 (23.9)
\$48 000-\$62 999	3139 (26.7)	891 (28.1)	16 953 (27.4)	2913 (27.2)	2596 (26.0)	1026 (26.4)
\$63 000 or more	3414 (29.1)	936 (29.6)	18 109 (29.3)	2748 (25.6)	3036 (30.5)	1130 (29.1)
Comorbidity with Charlson Deyo score (%)						
0	8867 (75.5)	2433 (76.8)	40 301 (65.2)	6660 (62.1)	8249 (82.8)	3208 (82.5)
1	1892 (16.1)	536 (16.9)	15 271 (24.7)	2952 (27.5)	1319 (13.2)	561 (14.4)
2	648 (5.5)	150 (4.7)	4621 (7.5)	834 (7.8)	268 (2.7)	76 (2.0)
3+	330 (2.8)	48 (1.5)	1638 (2.6)	272 (2.5)	132 (1.3)	44 (1.1)
Urban vs rural environment (%)						
Urban	9804 (83.5)	2644 (83.5)	51 229 (82.9)	8620 (80.4)	8655 (86.8)	3284 (84.4)



**TABLE 2** Summary of HR ranges and P-values among the combinations of studies run

Propensity score matching analysis	Landmark analysis and left truncation analysis	Lung			Prostate			Breast				
		N	HR range	% of all the P-values <.05	N	HR range	% of all the P-values <.05	N	HR range	% of all the P-values <.05		
No	No landmark analysis or truncation	72 549	(0.759,0.85)	100	14 904	(0.893,0.945)	100	24 805	13 857	(0.603,0.674)	100	100
	Left truncation only	72 411	(0.783,0.874)	100	14 865	(0.916,0.967)	43.152	0	13 835	(0.742,0.79)	100	100
	1 month landmark	29 904	(0.837,0.932)	100	9428	(0.997,1.051)	0	0	4180	(0.811,0.886)	85.632	0
	6 month landmark	48 505	(0.797,0.885)	100	12 872	(0.938,0.976)	5.017	0	9708	(0.77,0.846)	100	100
	12 month landmark	30 245	(0.749,0.813)	100	11 269	(0.892,0.922)	100	3.54	9876	(0.732,0.797)	100	100
Yes	No landmark analysis or truncation	21 436	(0.777,0.812)	100	6334	(0.943,1.015)	0	0	7778	(0.613,0.668)	100	100
	Left truncation only	21 422	(0.83,0.859)	100	6330	(0.824,0.894)	0	0	7774	(0.733,0.789)	100	0
	1 month landmark	7292	(0.839,0.876)	30.75	3298	(0.947,1.094)	0	0	722	(0.257,289,405)	1.803	0.684
	6 month landmark	15 606	(0.851,0.893)	100	5558	(0.979,1.059)	0	0	4370	(0.839,0.956)	0	0
	12 month landmark	10 392	(0.717,0.763)	100	4942	(1.113,1.298)	0.012	0	6216	(0.88,0.938)	0	0

Note: HR range refers to the range of hazard ratios (HRs) that we were able to generate among candidate models by using covariate combinations. The HR range can be visualized by the left-to-right spread of points in Figure 2. There are over 300 000 candidate models because of the different biostatistical methods applied, and the number of covariates within each model.

**TABLE 3** Summary of errors and correction methods in the current work

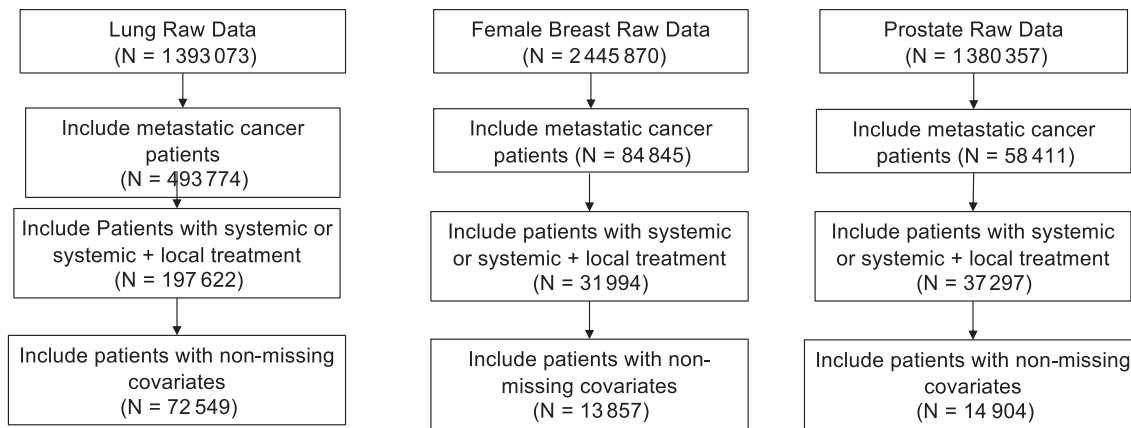
Type of error	Description	Correctional method used in our study	Tables/figures in our study	Overall effect
Selection bias <sup>16-18</sup>	Occurs when the baseline characteristics of patients who received systemic and local therapy differ from patients who received systemic therapy alone	<ul style="list-style-type: none"> <li>We applied propensity score matching approaches with the goal of simulating a randomized clinical trial by balancing the baseline characteristics of these two patient cohorts in balancing the distributions of the covariates between the two groups of patients to reduce the selection bias</li> </ul>	<ul style="list-style-type: none"> <li>Table 2 displays <i>P</i>-values and hazard ratio (HR) range with and without propensity score matching.</li> <li>Figure 2 displays volcano plots with and without propensity score matching.</li> </ul>	Propensity score matching balanced the two treatment groups and brought HRs closer to 1.
Immortal time bias <sup>7,19-21</sup>	Occurs when patients are at zero risk of death from time of diagnosis to time of treatment	<ul style="list-style-type: none"> <li>We define time to treatment as time from diagnosis to systemic or local therapy, whichever occurred first, regard time to treatment as left truncation time, and apply counting process style input</li> <li>We utilized the “landmark method” to adjust for immortal time bias by only including patients who survived after some fixed time point, labeled as “landmark”, into the data analysis</li> </ul>	<ul style="list-style-type: none"> <li>Table 2 displays <i>P</i>-values and HR range with and without left truncation and landmark analysis.</li> <li>Figure 2 displays volcano plots with and without left truncation and landmark analysis.</li> <li>Table S3 provides a schematic representation of left truncation and landmark analysis.</li> </ul>	Immortal time bias adjustment brought HRs closer to 1.
“Vibration of effects” due to covariate combination bias <sup>6</sup>	Occurs when unique sets of covariates leads to different results on multivariable analysis. Different researchers may select different covariates at their discretion; thus, the overall effects may vary.	<ul style="list-style-type: none"> <li>We conduct over 300 000 covariate combinations (all combination sets with the number of <math>\sum_{i=1}^p \binom{p}{i}</math> given <i>p</i> covariates) and plot the resulting vibrations of effects by comparing the HR and <i>P</i>-value each one produce.</li> </ul>	<ul style="list-style-type: none"> <li>Figure 1 displays patients excluded from study due to missing covariates.</li> <li>Figure 2 uses volcano plots to display <i>P</i>-value vs hazard ratio, resulting from all combinations. Each dot on a volcano plot represents a different study with a unique covariate combination. For example, one dot may control for race and age, and another dot may only control for race.</li> <li>Table S2 lists the covariates used in this article, along with their original variable name in the database.</li> </ul>	Covariate combination adjustment produced hazard ratios on both sides of 1, or equaling 1. With additional selection bias adjustment and immortal time bias adjustment, the survival impact of an intervention could be magnified or nullified.

breast cancer, 9968 received systemic and 3889 received systemic + local therapy. Table 1 shows patient characteristics. Patients in the systemic + local therapy group were more likely to be white, have private insurance, have NO disease, come from a location with a more educated and more wealthy population and have fewer comorbidities.

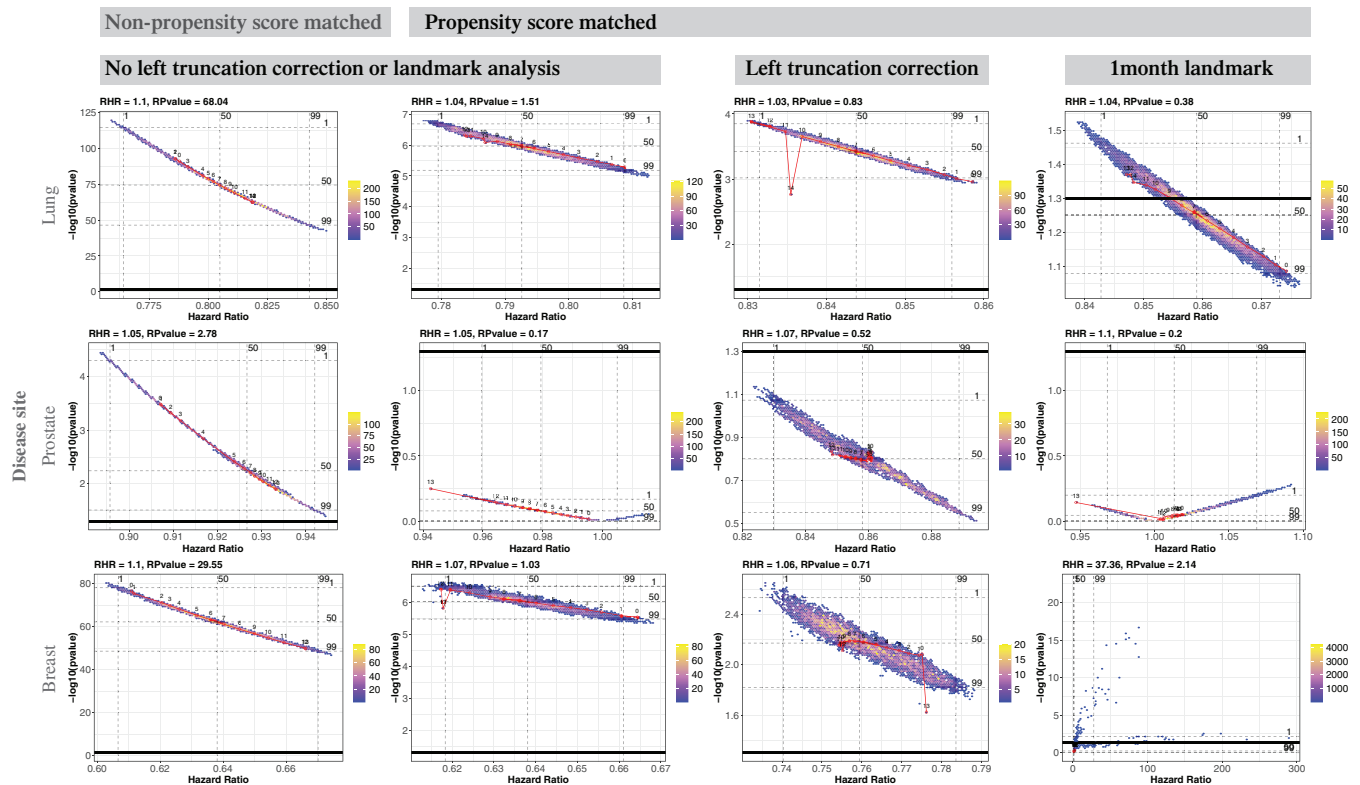
Table 2 shows summaries of the HR ranges with the different biostatistical methods, using >300 000 regression models (each model representing a publishable study). In the most basic analysis, without propensity score matching (top panels), left truncation adjustment or landmark analysis, with combinations of covariates, all of the HRs were <1 (generally, 0.60-0.95, favoring addition of local therapy), with all *P*-values <.001. The use of a left truncation adjustment had minimal

impact on the reported HRs for female breast and lung cancer, but made the HRs not statistically significant for prostate cancer. The addition of landmark analysis similarly had a greater effect on the reported HRs for prostate cancer, but less so for female breast and lung cancer. For all three cancers, the use of more statistical methods to reduce the selection bias caused reported HR ranges to approach 1.0.

With propensity score matching (Table 2, lower panels), but without landmark analysis or left truncation adjustment, all HRs were <1 and *P*-values were <.001 for lung and breast cancer, but not for prostate cancer. The addition of left truncation adjustment produced HRs >1. The combination of propensity score matching, left truncation



**FIGURE 1** Patient selection flow diagram. We included patients with metastatic prostate, breast and lung cancer. Within each cohort of patients, two groups were created: one that received systemic therapy alone, and one that received systemic therapy + local therapy (typically surgery and/or radiotherapy). Detailed selection criteria are presented in the Supporting Information File S1



**FIGURE 2** Volcano plots of the  $-\log(P\text{-value})$  vs hazard ratio (HR). These plots show variability of effect sizes by modifying biostatistical methods. Each point on the subplots represents a publishable comparative effectiveness study that evaluates addition of local therapy to systemic therapy for metastatic cancer. There are over 300 000 studies plotted. HRs and  $-\log_{10}(P\text{-value})$  obtained from different combinations of adjustments are used to visualize the vibrations of effects with volcano plots.<sup>6</sup> HRs  $<1$  suggest benefit with systemic + local therapy and HRs  $>1$  suggest benefit with systemic therapy alone. The bold black horizontal line represents significance of  $P\text{-value} = .05$ , and points plotted above this line have a  $P\text{-value} < .05$ . Red circles with a specified number  $k$  highlight the median HR and  $P\text{-value}$  in the models with a specified number of adjustment variables ( $k$ ). The dotted lines depict the 1st, median and 99th percentile of the X and Y axes, respectively. In the upper left notes of each subplot, we also compute the summary statistics, including the relative hazard ratio (RHR, calculated as 99% quantile of HR/1% quantile of HR) and relative  $P\text{-value}$  (RPvalue, calculated as 1% quantile of  $[-\log_{10}(P\text{-value})]$  – 99% quantile of  $[-\log_{10}(P\text{-value})]$ ). For simplicity, in the landmark analysis, only the data from the 1-month landmark are presented; other landmark analyses were performed, yielding results shown in Table 2. In summary, we generated contrary outcomes, with HRs on both sides of 1, and 100 log-fold differences in  $P\text{-values}$ . Our results suggest that some retrospective observational studies may find a treatment helps, and another may find it does not, simply based on analytic choices [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



adjustment, landmark analysis and covariate combinations generally produced *P-values* that were no longer significant at the .001 threshold and sometimes the .05 threshold. Further details about patients and the impact of biostatistical adjustment on the HRs are in Supporting Information File S2. In summary, by modifying biostatistical techniques, we were able to generate contrary outcomes in HRs and 100-fold differences in  $-\log_{10}(P\text{-values})$ . Details of statistical analysis combinations are presented in Supporting Information File S4.

Table 3 provides a summary of the principal sources of error/bias that are present in retrospective comparative effectiveness studies, the correctional methods used in our work, the related tables and figures and the overall effect of these corrections. In summary, there were three sources of bias that we attempted to mitigate: selection bias,<sup>16-18</sup> immortal time bias<sup>7,19-21</sup> and “vibration of effects” due to covariate combination bias.<sup>6</sup> When adjusting for selection bias and immortal time bias, we found that the HRs were brought closer to 1. Covariate combinations produced HRs on both sides of 1, or equaling 1. In summary, by modifying biostatistical techniques, we were able to generate contrary outcomes in HRs and 100-fold differences in  $-\log_{10}(P\text{-values})$ .

## 4 | DISCUSSION

Prior works have shown discordance between randomized controlled trials and observational studies. Our results show why this discordance exists. We used an example of treatment of the primary tumor in the setting of metastatic disease in three cancer sites to demonstrate that, by modifying parts of the biostatistical model, the resulting HR could favor either arm, or show equivalence between two treatment paradigms, and have 100 log-fold differences in *P-values*. Our analysis suggests that conclusions drawn from observational studies are highly sensitive to the choice of analytic approach. The more advanced/appropriate the statistical adjustment techniques are (as long as there is not a certain level of adjustable indication bias), the more one will favor the null hypothesis. When a controversial topic is being studied by multiple independent investigators, each pursuing different analytic plans, a variety of outcomes may be reached. These data support a call for authors to supply “a priori” analytic plans/protocols, and for journals/reviewers to demand these protocols and preferably prespecified sensitivity analyses as appropriate to determine robustness of results.

Comparative effectiveness research is a core component of evaluating competing treatment options in medicine. In the 1900s, the advent of randomized controlled trials created the “gold standard” for comparative effectiveness research.<sup>2,3</sup> Randomized trials have several benefits over retrospective studies. For example, combinations of covariates used in different retrospective studies attempting to address the same question result in variability in reported estimates and effect sizes; this phenomenon has been shown in dietary intervention studies, and the current work is the first to show it in treatment interventions in oncology.<sup>6</sup> Additionally, retrospective studies have been faulted for immortal time bias,<sup>7</sup> wherein a cohort of patients receiving a more aggressive treatment (eg, surgery +

radiotherapy vs surgery alone) is guaranteed to be alive longer than the comparison group because it must live long enough to receive the additional treatment.

The variability seen in our study would likely only increase with the addition of other factors. Since ours is an observational study, we are unable to evaluate unknown confounders on outcomes, which is one benefit of randomization.<sup>22,23</sup> Other covariates could be included in our vibration of effects permutations (eg, particular comorbidities, performance status, chemotherapy doses). This analysis uses data from cancer patients in the United States alone, and future studies may probe the stability of conclusions in other datasets.

We performed this analysis for specific disease sites in oncology, and our results may not apply to other comparative effectiveness research questions (eg, surgery vs radiotherapy; drug A vs drug B). However, we chose the disease sites and scenarios for several reasons. (a) Prostate, breast and lung cancer are common in developed countries<sup>24</sup>; thus, the results would be generalizable. (b) Metastatic prostate and breast cancer are relatively indolent while lung cancer is more aggressive; since we saw signal of biostatistical uncertainty in all three, we do not believe that lethality of disease would justify using retrospective data for comparative effectiveness research. (c) Limited prospective studies on the value of local control have been published for these cancers. For example, a 2015 randomized trial showed local control to have no impact on survival in breast cancer,<sup>10</sup> but a retrospective analysis and meta-analysis published in 2020 showed the opposite.<sup>25,26</sup> Similarly, for prostate cancer, a 2016 retrospective database analysis using propensity score matching showed HR of 0.62 with local radiotherapy.<sup>12</sup> From 2017 to 2019, several randomized trials of men with metastatic prostate showed limited benefit of local control.<sup>11,27</sup> For lung cancer, prospective data are sparse, and we suspect that upcoming randomized trials will show minimal impact of local control.

Although retrospective data may not be ideal for comparative effectiveness research, they are ideal at showing “real world” outcomes after treatment, studying epidemiology and evaluating quality of care. Retrospective data may also be hypothesis-generating for prospective studies, particularly if they show a large effect size. Some comparative effectiveness research articles can still be very useful if robustly designed/analyzed, even if not aligning perfectly with randomized trial data (given that the populations may be quite different). In addition, one national database (eg, the NCDB) does not represent all observational databases and the conclusions are still quite extreme and potentially biased to fit a narrative that observational comparative effectiveness research is irredeemably biased and manipulated. Thus, there retrospective and prospective analyses are complementary.

## 5 | CONCLUSION

Our analysis suggests that conclusions drawn from observational studies are highly sensitive to the choice of analytic approach. When a controversial topic is being studied by multiple independent investigators, each pursuing different analytic plans, a variety of outcomes



may be reached. In order to mitigate this tendency toward bias in comparative effectiveness research, journals should demand prespecified analytic plans and protocols from authors as needed to determine the robustness of results.

### CONFLICT OF INTEREST

Vinay K. Prasad is supported by a research funding grant from Arnold Ventures; royalties from Johns Hopkins Press, Medscape and MedPage; consulting fees from UnitedHealthcare; speaking fees from Evicore and New Century Health; and Patreon funding from the Plenary Session podcast. This support is not relevant to the subject of the research but is included with the goal of complete transparency. The other authors have no conflict of interest to declare.

### DATA AVAILABILITY STATEMENT

The data are available by request of an investigator to the National Cancer Database, at the American College of Surgeons and American Cancer Society: <https://www.facs.org/qualityprograms/cancer/ncdb>. Further information is available from the corresponding author upon request.

### ETHICS STATEMENT

The study was approved by the Institutional Review Board. It uses previously collected, de-identified data from a national registry. The informed consent was waived.

### ORCID

Nicholas G. Zaorsky  <https://orcid.org/0000-0002-4932-1986>

### REFERENCES

- Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *Evid Based Med*. 2016;21:125-127.
- Meldrum ML. A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. *Hematol Oncol Clin North Am*. 2000;14(745-60):vii-760.
- Bothwell LE, Podolsky SH. The emergence of the randomized. *Controlled Trial N Engl J Med*. 2016;375:501-504.
- Soni PD, Hartman HE, Dess RT, et al. Comparison of population-based observational studies with randomized trials in oncology. *J Clin Oncol*. 2019;37:1209-1216.
- Kumar A, Guss ZD, Courtney PT, et al. Evaluation of the use of cancer registry data for comparative effectiveness research. *JAMA Netw Open*. 2020;3:e2011985.
- Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol*. 2015;68:1046-1058.
- Park HS, Gross CP, Makarov DV, Yu JB. Immortal time bias: a frequently unrecognized threat to validity in the evaluation of postoperative radiotherapy. *Int J Radiat Oncol Biol Phys*. 2012;83:1365-1373.
- Silberzahn R, Uhlmann EL, Martin DP, et al. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv Methods Pract Psychol Sci*. 2018;1:337-356.
- Tannock IF. Removing the primary tumor after the cancer has spread. *N Engl J Med*. 2001;345:1699-1700.
- Badwe R, Hawaldar R, Nair N, et al. Locoregional treatment versus no treatment of the primary tumour in metastatic breast cancer: an open-label randomised controlled trial. *Lancet Oncol*. 2015;16:1380-1388.
- Parker CC, James ND, Brawley CD, et al. Radiotherapy to the primary tumour for newly diagnosed, metastatic prostate cancer (STAMPEDE): a randomised controlled phase 3 trial. *Lancet*. 2018;392:2353-2366.
- Rusthoven CG, Jones BL, Flaig TW, et al. Improved survival with prostate radiation in addition to androgen deprivation therapy for men with newly diagnosed metastatic prostate cancer. *J Clin Oncol*. 2016;34:2835-2842.
- Billing PS, Miller DL, Allen MS, Deschamps C, Trastek VF, Pairolero PC. Surgical treatment of primary lung cancer with synchronous brain metastases. *J Thorac Cardiovasc Surg*. 2001;122:548-553.
- Bilimoria KY, Stewart AK, Winchester DP, Ko CY. The National Cancer Data Base: a powerful initiative to improve cancer care in the United States. *Ann Surg Oncol*. 2008;15:683-690.
- Boffa DJ, Rosen JE, Mallin K, et al. Using the National Cancer Database for outcomes research: a review. *JAMA Oncol*. 2017;3:1722-1728.
- Austin PC. An Introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399-424.
- Lalani N, Jimenez RB, Yeap B. Understanding propensity score analyses. *Int J Radiat Oncol Biol Phys*. 2020;107:404-407.
- Soni PD. Selection bias in population registry-based comparative effectiveness research. *Int J Radiat Oncol Biol Phys*. 2019;103:1058-1060.
- Suissa S. Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol*. 2008;167:492-499.
- Cain KC, Harlow SD, Little RJ, et al. Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. *Am J Epidemiol*. 2011;173:1078-1084.
- Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol*. 1983;1:710-719.
- Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 1998;54:948-963.
- Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B*. 1983;45:212-218.
- Fitzmaurice C, Dicker D, Pain A, et al. The global burden of cancer 2013. *JAMA Oncol*. 2015;1:505-527.
- Gera R, Chehade H, Wazir U, et al. Locoregional therapy of the primary tumour in de novo stage IV breast cancer in 216 066 patients: a meta-analysis. *Sci Rep*. 2020;10:2952.
- Pons-Tostivint E, Kirova Y, Lusque A, et al. Radiation therapy to the primary tumor for de novo metastatic breast cancer and overall survival in a retrospective multicenter cohort analysis. *Radiother Oncol*. 2020;145:109-116.
- Boeve LMS, Hulshof M, Vis AN, et al. Effect on survival of androgen deprivation therapy alone compared to androgen deprivation therapy combined with concurrent radiation therapy to the prostate in patients with primary bone metastatic prostate cancer in a prospective randomised clinical trial: data from the HORRAD trial. *Eur Urol*. 2019;75:410-418.

### SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Zaorsky NG, Wang X, Lehrer EJ, et al. Retrospective comparative effectiveness research: Will changing the analytical methods change the results? *Int. J. Cancer*. 2022;150(12):1933-1940. doi:10.1002/ijc.33946