original report

Comparison of Population-Based Observational Studies With Randomized Trials in Oncology

Payal D. Soni, MD[1]; Holly E. Hartman, MS[2]; Robert T. Dess, MD[2]; Ahmed Abugharib, MD[3]; Steven G. Allen, PhD[2]; Felix Y. Feng, MD[4]; Anthony L. Zietman, MD[5]; Reshma Jagsi, MD, DPhil[2]; Matthew J. Schipper, PhD[2]; and Daniel E. Spratt, MD[2]

abstract

PURPOSE Comparative efficacy research performed using population registries can be subject to significant bias. There is an absence of objective data demonstrating factors that can sufficiently reduce bias and provide accurate results.

METHODS MEDLINE was searched from January 2000 to October 2016 for observational studies comparing two treatment regimens for any diagnosis of cancer, using SEER, SEER-Medicare, or the National Cancer Database. Reporting quality and statistical methods were assessed using components of the STROBE criteria. Randomized trials comparing the same treatment regimens were identified. Primary outcome was correlation between survival hazard ratio (HR) estimates provided by the observational studies and randomized trials. Secondary outcomes included agreement between matched pairs and predictors of agreement.

RESULTS Of 3,657 studies reviewed, 350 treatment comparisons met eligibility criteria and were matched to 121 randomized trials. There was no significant correlation between the HR estimates reported by observational studies and randomized trials (concordance correlation coefficient, 0.083; 95% CI, −0.068 to 0.230). Forty percent of matched studies were in agreement regarding treatment effects (κ, 0.037; 95% CI, −0.027 to 0.01), and 62% of the observational study HRs fell within the 95% CIs of the randomized trials. Cancer type, data source, reporting quality, adjustment for age, stage, or comorbidities, use of propensity weighting, instrumental variable or sensitivity analysis, and well-matched study population did not predict agreement.

CONCLUSION We were unable to identify any modifiable factor present in population-based observational studies that improved agreement with randomized trials. There was no agreement beyond what is expected by chance, regardless of reporting quality or statistical rigor of the observational study. Future work is needed to identify reliable methods for conducting population-based comparative efficacy research.

J Clin Oncol 37.

INTRODUCTION

In the last decade, there have been significant strides in cancer treatment, with the introduction of precision medicine, new targeted therapies and immune modulating agents, and technologic advances in surgery, imaging, and radiotherapy. As the field of oncology continues to evolve, comparative efficacy research (CER) is paramount in understanding how new therapies should be integrated into patient care and in developing health care policies. Randomized controlled trials (RCTs) are the gold standard for comparing treatment efficacy. However, because of the financial burden and time of running randomized trials and the selectivity of patients eligible for most randomized trials, there is growing interest in alternative CER methods that can keep up with the pace of change in oncology.

Population-based observational studies have been increasingly leveraged to compare treatment outcomes and have been used to influence clinical decisions and guideline recommendations.[1-6] However, without randomization, associations between treatment and outcomes suggested by observational research may be influenced by confounding factors. Furthermore, incomplete or incorrect data in registries may bias outcomes.[7,8] Several national organizations, including the Agency for Healthcare Research and Quality,[9] the Institute of Medicine,[10] and the American Society of Clinical Oncology,[11] have endorsed the use of observational CER to complement RCTs, with the caveat that it must be performed using rigorous methodologies to minimize bias. However, no study to our knowledge has objectively demonstrated methods that reproducibly and sufficiently reduce bias in population-based CER.

Herein, we perform a comprehensive analysis of modern population-based comparative effectiveness studies focused on cancer treatments and compare

ASSOCIATED CONTENT

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on February 26, 2019 and published at jco.org on March 21, 2019: DOI https://doi.org/10.1200/JCO.18.01074

the results with corresponding RCTs. We hypothesized that there would be overall poor agreement between the two study types but that factors that improve agreement could be identified.

METHODS

Study Selection

A systematic MEDLINE search was performed using controlled vocabulary (Data Supplement) to identify all observational studies published between January 2000 and October 2016 using the three most common population-based registries for cancer research in the United States: SEER, SEER-Medicare, and the National Cancer Database (NCDB). These three databases are described in the Data Supplement.[12] Search results were screened to identify treatment comparisons for any cancer with an overall (OS) or cancer-specific survival end point. When studies included multiple comparisons, each comparison was recorded. Comparisons were excluded if sample size was not provided.

Identifying Matching RCTs

For every observational comparison, MEDLINE was searched for matching RCTs on the basis of treatment comparison, study population, and presence of a survival end point. This search was supplemented by the observational study discussion, National Comprehensive Cancer Network guidelines, and published reviews. If multiple RCTs were identified, they were additionally screened for the best-matching inclusion criteria by age, stage of disease, and treatment details. Finally, if multiple studies remained, the largest study was selected. To quantify how well matched an observational study and RCT were, a match level was defined on the basis of the age and stage inclusion criteria used in both studies. The match for each criterion was scored on a scale of 0 to 2, indicating different, overlapping, or exactly the same criteria. These were added for a total match level between 0 and 4, in which 1 to 2 indicated a moderately matched pair and 3 to 4 indicated a well-matched pair.

Data Extraction

Two independent investigators (P.D.S. and D.E.S.) extracted data from each study. Treatment comparisons were grouped into four categories: addition of surgery, radiotherapy, systemic therapy, or other (Data Supplement). Data on publication year, journal impact factor, disease site, treatment comparison, sample size, disease stage, and age of patients were collected. Quality of reporting was captured by whether a study reported age (eg, mean, median, or range) of included patients, median follow-up, and extent of and methods for handling missing data. Statistical rigor was captured by whether adjustments for age, extent of disease, comorbidities, and geographic region were performed. Use of advanced statistical methods was captured, defined as use of multivariable models, propensity score methods, instrumental variable approaches, and sensitivity analyses. Finally, survival data, hazard ratios (HRs), and 95% CIs were extracted.

Statistical Analysis

The primary aim of this study was to characterize agreement between population-based observational studies and matching RCTs in terms of treatment effect estimates on OS or cancer-specific survival. Qualitatively, agreement was determined if both studies reported a statistically significant OS or cancer-specific survival benefit with the same treatment or both reported no significant difference. Significance was defined by each study. Weighted and unweighted κ statistics were calculated to quantify and test for significant agreement beyond chance alone.

Quantitatively, agreement was assessed by HR estimates. A concordance correlation coefficient (CCC) was calculated to quantify concordance between the HRs reported in matched comparisons. This measure captures deviation between each point and the line of perfect agreement ($y = x$).[13] To account for potential correlation induced by matching, the observational HR was modeled as a function of the RCT HR in a mixed effect model including a random intercept for each RCT and observational study. Furthermore, the percentage of observational HRs falling within the 95% CIs reported in the RCTs was calculated.

In the primary analyses, each pair was weighted equally. In sensitivity analyses, different weighting schemes were used to give each RCT the same cumulative weight (to account for RCTs matched to multiple observational comparisons) and to give each observational study the same weight (to account for observational studies with multiple treatment comparisons). Information on methodology is detailed in the Data Supplement. Furthermore, both quantitative and qualitative analyses of agreement were repeated for a subset of rigorously performed observational studies that were well matched to randomized trials (match level, 3 to 4). All observational studies included in this subset adjusted for age and extent of disease and performed at least one advanced statistical strategy (multivariable models, propensity score methods, instrumental variable approaches, or sensitivity analyses).

Univariable and multivariable logistic regression models were fit to identify predictors of qualitative agreement. Variables included year of publication, journal impact factor, data source, sample size, disease type, treatment modality, observational outcome, primary end point of RCT, and measures of reporting quality, statistical rigor, and match level. A Cochrane-Armitage trend test was performed to determine if a better match was associated with better agreement. Statistical significance for the univariable model was set at $P < .0016$. This P value was derived using a Bonferroni correction to account for multiple testing. For all other statistical analyses, $P < .05$ was considered statistically significant.

RESULTS

A total of 3,657 observational studies were identified from the initial search. After applying eligibility criteria, 456 were selected, of which 172 (38%) included multiple treatment comparisons. In total, 755 comparisons were identified, and 350 (46.4%) were matched to 121 RCTs (Fig 1). Multiple observational comparisons (median, two) investigating the same treatments in similar populations were matched to the same RCT (Data Supplement).

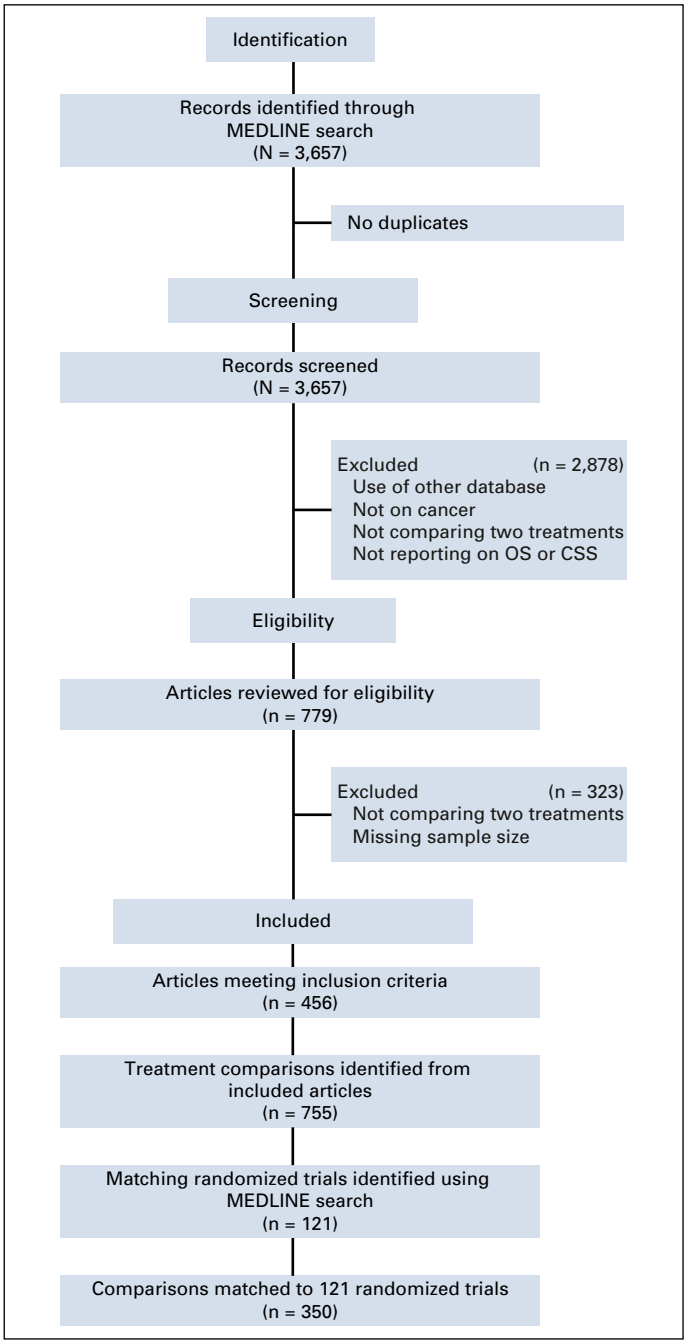

FIG 1. Preferred reporting items for systematic reviews and meta-analyses. CSS, cancer-specific survival; OS, overall survival.

Characteristics of all observational comparisons and the matched subset are listed in Table 1, demonstrating that the quality of the matched subset is representative of all observational comparisons. RCT characteristics are provided in the Data Supplement. Comparative observational studies have been increasing in publication over the last 17 years, with 50% published in the last 4 years (Data Supplement). Multivariable analyses (81%) were regularly used. Propensity adjustments (27%), instrumental variable analyses (4%), and sensitivity analyses (12%) were infrequent.

A majority (68%) of published observational comparisons reported a statistically significant survival difference between compared treatments. However, only 31% of the RCTs reported significant survival differences (Data Supplement). Observational studies were significantly more likely to demonstrate longer survival with the addition of surgery than with radiotherapy (odds ratio [OR], 0.33; $P < .001$) or systemic therapy (OR, 0.4; $P = .003$). No such pattern was identified among RCTs (Data Supplement).

Only 40% (n = 138) of matched pairs reported the same conclusion. The κ statistic for agreement was 0.037 (95% CI, −0.027 to 0.01), suggesting no agreement beyond that expected by chance alone. Sensitivity analyses provided similar results (Data Supplement). In 91% of the disagreeing pairs, one study reported a significant survival difference between treatments, whereas the other reported no significant survival difference. In the remaining 9% (n = 20 pairs), both studies found significant survival differences but in opposite directions.

HRs for OS were provided by both studies in 165 matched pairs (47%). No significant correlation was identified between the matched HR estimates, with a CCC of 0.083 (95% CI, −0.068 to 0.230; Data Supplement). Additional analyses of correlation by treatment modality are provided in the Data Supplement. Mixed effect regression models demonstrated that the RCT HR was not a statistically significant predictor for the observational HR (estimate, 0.093; 95% CI, −0.16 to 0.345; $P = .47$). Sensitivity analyses described in the Data Supplement similarly failed to demonstrate a relationship. In 151 pairs, the RCTs also provided 95% CIs for their HRs. Only 62% of the observational HRs fell within the 95% CIs reported by the RCTs (Data Supplement).

Analysis of Rigorous Well-Matched Observational Studies

Of the 350 matched observational studies, 238 (68%) met our criteria for being rigorous and well matched. Agreement among matched pairs remained low at 40.3% (n = 96). Of those pairs that agreed, 40.6% (n = 39) had significant findings and 59.4% (n = 57) did not. Of those that disagreed, 5.6% (n = 8) had significant findings in opposite directions and 94.4% (n = 134) had differing results.

TABLE 1. Characteristics of Observational Studies

Characteristic	No. (%) of Observational Studies		
	All (n = 755)	Matched (n = 350)	*P*
Sample size			.22
Mean	8,686	10,610	
SD	26,017	23,159	
Median	2,096	3,430	
IQR	807-6,612	1,306-9,084	
Impact factor			.02
Mean	5.3	6.3	
SD	5.6	7.4	
Median	4.1	4.3	
IQR	3.0-5.7	3.0-5.7	
Year published			.33
≤ 2007	89 (12)	51 (15)	
2008-2012	270 (36)	129 (37)	
2013-2016	396 (53)	170 (49)	
Data source			.55
SEER	473 (63)	213 (61)	
SEER-Medicare	169 (22)	89 (25)	
NCDB	111 (15)	46 (13)	
Multiple	2 (0.3)	2 (1)	
Disease type			< .001
Breast	77 (10)	44 (13)	
Endocrine	31 (4)	1 (0.3)	
CNS	36 (5)	7 (2)	
GI	208 (28)	108 (31)	
Genitourinary	101 (13)	50 (14)	
Gynecologic	57 (8)	28 (8)	
Head and neck	47 (6)	12 (3)	
Hematologic	45 (6)	11 (3)	
Pediatric	8 (1)	1 (0.3)	
Sarcoma	30 (4)	11 (3)	
Skin	6 (1)	2 (1)	
Thoracic	109 (14)	75 (21)	
Type of comparison			.002
Addition of surgery	128 (17)	34 (10)	
Addition of radiotherapy	296 (39)	163 (47)	
Addition of systemic therapy	106 (30)	60 (17)	
Other	294 (39)	122 (35)	
Reporting quality			
Any age metric reported	567 (75)	267 (76)	.78
Median follow-up reported	290 (38)	138 (40)	.80
Extent of missing data reported	548 (73)	225 (71)	.54
Handling of missing data reported	475 (63)	204 (64)	.80
Statistical rigor			
Adjustments			
Age	652 (86)	312 (89)	.23
Extent of disease	654 (87)	318 (91)	.06
Comorbidities	256 (34)	124 (35)	.67
Geographic region	259 (34)	135 (39)	.19
Advanced statistical methods			
Multivariable analysis	611 (81)	292 (83)	.36
Propensity adjustment	202 (27)	124 (35)	.004
Instrumental variable	27 (4)	16 (5)	.53
Sensitivity analysis	87 (12)	50 (14)	.23

Abbreviations: IQR, interquartile range; NCDB, National Cancer Database; SD, standard deviation.

There were 121 pairs within this subset in which both studies reported an HR. Concordance among these HRs remained low (CCC, 0.128; 95% CI, 0.046 to 0.295; Fig 2A). The RCT HR was again not a statistically significant predictor of the observational HR (estimate, 0.108; 95% CI, −0.038 to 0.254). Only 64% (n = 75) of observational HRs from this subset fell within the 95% CIs reported by the well-matched RCTs (Fig 2B).

Predictors of Agreement

There was no significant improvement in agreement between observational and matched RCTs among studies that used adjustments or advanced statistical strategies to address bias (*P* = .20 to .94; Fig 3). There was also no significant association between how well matched the study populations were and agreement (*P* = .63; Data Supplement). On univariable analysis, only the conclusion of the

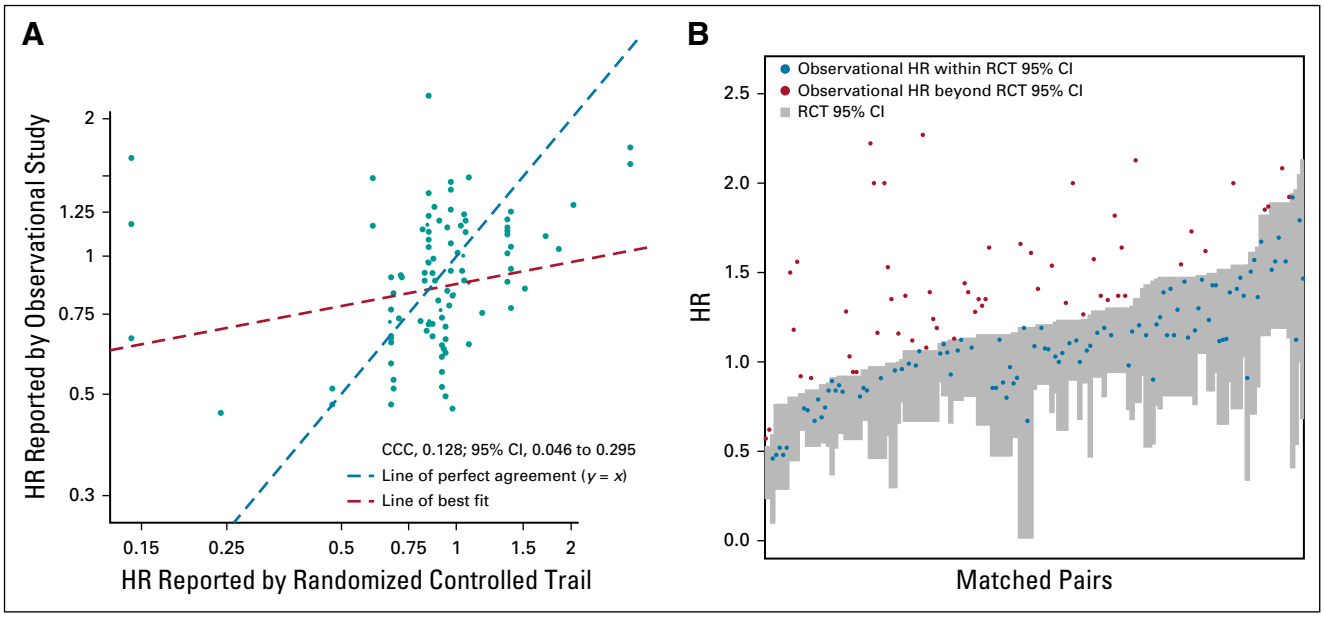

FIG 2. Comparison of hazard ratios (HRs) reported by rigorously performed, well-matched observational studies and randomized trials. (A) Scatter plot of HR reported by observational study versus randomized controlled trial (RCT) for each matched pair (n = 121). *x*- and *y*-axes presented on log scale. Red dashed line represents the line of best fit; teal dashed line represents where the line of best fit would be if the HRs from the observational study and RCT were equal. (B) RCT HR 95% CI (gray boxes) with observational study HR estimates (red and blue dots). Matched pairs ordered by the upper CI limit of RCT. HRs were inverted as necessary to ensure that both HRs were reported relative to the same reference treatment and that the observational HR was greater than the randomized trial HR. HR < 1 indicates improved survival with the comparator treatment compared with the reference. CCC, concordance correlation coefficient.

observational study predicted agreement. Observational studies demonstrating no significant survival difference were much more likely to agree with corresponding RCTs (OR, 11.64; $P < .001$; Data Supplement). On multivariable analysis, after adjusting for disease type and stage, the conclusion of the observational study remained the strongest predictor of agreement (OR, 33.16; $P < .001$). Studies assessing the addition of radiotherapy (OR, 0.19; $P < .001$) or studies in the treatment comparison group termed other (OR, 0.1; $P < .001$) had a low likelihood of agreeing with their RCTs. The area under the curve for this model was 0.89 (Data Supplement). Registry-based studies typically have large sample sizes and may be able to identify small but statistically significant differences in survival. Of the 211 disagreeing observational studies, 169 provided a HR (Data Supplement). Only 12% of these studies reported a HR between 0.9 and 1.1 and 34% between 0.8 and 1.2, suggesting this phenomenon alone does not explain the lack of agreement

There were 70 rigorous observational studies that were well matched to RCTs with a primary end point of OS. In 35 of these pairs, the observational study was positive and the RCT reported either no difference or the opposite finding. To determine if insufficient sample size contributed to these differences, individual patient-level meta-analyses were identified and matched to seven of these observational studies. Among these seven pairs, only two meta-analyses (28.6%) agreed with the observational studies. Three meta-analyses (42.9%) continued to demonstrate no significant difference between treatments, and two (28.6%) demonstrated better survival for the opposite treatment approach (Data Supplement).

DISCUSSION

Population-based comparative effectiveness studies have been published at exponentially increasing rates in oncology over the last 17 years, despite a lack of evidence supporting the reliability of this research method. To better understand the reliability of population-based CER, we characterized the agreement between population-registry studies and RCTs, in what we believe is the most comprehensive analysis of population-based CER in oncology. We found no agreement between population-based studies and RCTs regarding treatment effects beyond what would be expected by chance alone, even when we limited our analysis to statistically rigorous population-based studies. This held true even in a more loosely defined metric of agreement assessing the HR estimates of the observational studies and the 95% CIs reported by the RCTs. Although we investigated several rationales for disagreement, including poor reporting quality, lack of statistical rigor, and differing study populations and sample sizes, we failed to identify a single modifiable factor that would reliably improve agreement.

There are several possible explanations for lack of agreement between observational studies and RCTs. Although

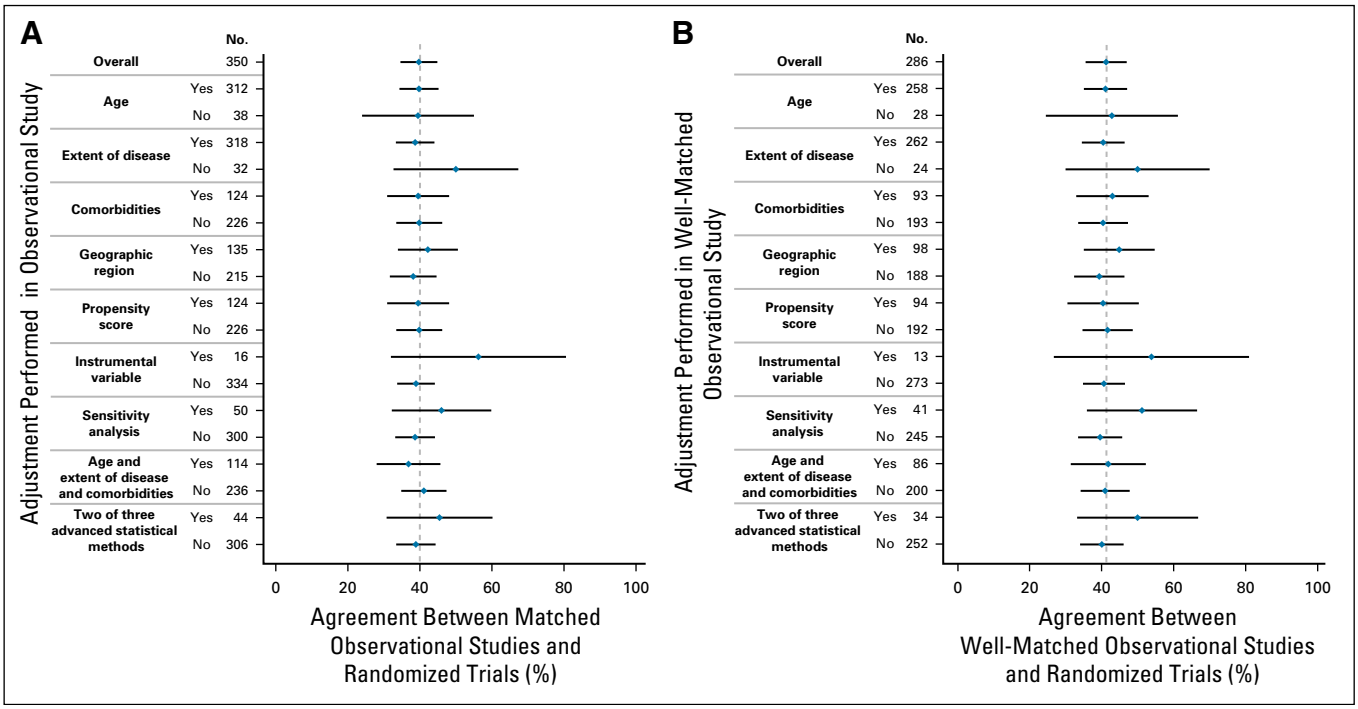

FIG 3. Agreement between observational studies and randomized controlled trials by adjustments performed in the observational study. (A) All matched observational studies and randomized trials (n = 350). (B) Well-matched observational studies and randomized trials defined by match level of 3 to 4 (n = 196). Gray line indicates match percentage for overall group.

RCTs have reduced bias and confounding compared with retrospective population-based CER, bias does exist within RCTs, and thus, RCT results do not inherently imply truth. RCTs can also be subject to flaws in design and analysis,[14] and it is unknown whether RCT results can be extrapolated beyond the exact population studied. However, well-done RCTs remain the least biased form and gold standard of CER. Furthermore, a comprehensive comparison of patients enrolled in SWOG trials and nontrial patients has suggested that the biases introduced into RCTs from stringent enrollment criteria only affect short-term outcomes. After the first year of follow-up, OS for patients enrolled in the standard arm of randomized trials is comparable to that of nontrial patients, suggesting generalizability of RCTs.[15] At a minimum, well-performed population-based CER studies should be able to generally replicate results from RCTs, when appropriately matched and adjusted, especially in cases where there are multiple RCTs demonstrating a consistent treatment effect. If this is not reliably possible, this raises concern for the validity of population-based CER studies, especially when combined with known limitations of the registries used.[8] This could result from limitations in the data accuracy of registries, data elements available for adjustment, or conduct of the analysis by the study authors. Although registries are becoming more robust with the availability of information regarding comorbidities and detailed staging and pathologic information with select tumor profiling, they still lack robust information on treatment intent, quality and compliance, duration of therapy, and subsequent therapy, as well as patients' functional status, smoking status, and overall health, which can affect treatment selection and survival independent of age and comorbidities.[16-18] Registries are also unable to capture the complexities of patient-physician interactions and preferences, which can alter a patient's treatment course.

Several authors have illustrated the effects of selection biases on observational research.[5,19] For example, in a SEER-Medicare analysis of men with prostate cancer, the addition of androgen deprivation therapy to radiotherapy resulted in an increased prostate cancer mortality, despite adjustment for stage, comorbidities, and propensity scores among other confounders. These findings directly contradict four RCTs that have consistently showed the exact opposite result.[20] Likewise, in a comparison between observation and active treatment for prostate cancer, SEER-Medicare data suggested patients undergoing surgery for prostate cancer, a treatment usually reserved for younger, healthier patients and those with better performance status, had better survival compared with a matched population of patients without cancer,[5] an unrealistic finding. This is largely consistent with our study in which observational comparisons investigating the benefit of surgery were significantly more likely to be positive than radiotherapy or systemic therapy comparisons. Our findings suggest that these biases may be pervasive throughout population-based studies in oncology.

We identified a greater than two-fold discrepancy in the percentage of observational and randomized studies reporting statistically significant survival differences between compared treatments. Although large observational studies may be able to identify statistical significance in small effect sizes, given the large sample sizes, fewer than 35% of disagreeing comparisons in our study had small effect sizes. These findings may instead reflect a publication bias.

To our knowledge, there have been only two smaller systematic comparisons of observational studies and RCTs.[1,2] Although both analyses found reasonable agreement, both were performed in an older era and involved a limited number of studies. Benson et al[1] analyzed comparative studies published between 1985 and 1998 in 120 journals, which excluded oncology journals. They identified only 19 different treatment comparisons. Concato et al[2] focused their search on meta-analyses published between 1991 and 1995 in five journals, which included 99 randomized and observational studies on five different clinical topics. Fewer than 15% of the observational studies in both reviews were performed using population-based registries, and fewer than 20% of the comparisons in both reviews reported a survival end point. A strength of our study is our comprehensive search method. By including all comparative efficacy studies published using the three databases of interest, we minimized biases that could have been introduced by study selection. Furthermore, our study focused entirely on population-based observational studies with survival end points, which are susceptible to more forms of bias than shorter-term end points. This is important, because intermediate end points may be more reliable and less subject to bias but are currently not available in SEER or NCDB registries.

Limitations of our study exist. We analyzed studies in oncology performed using the three major cancer databases in the United States. Conclusions cannot be made on studies performed in other disciplines or using other databases. Notably, we searched six additional US claims data sets, five Nordic cancer registries, and the Ontario Cancer Registry; there were only 29 eligible studies across these 12 registries over a 17-year period, compared with the more than 750 we identified using SEER, SEER-Medicare, and NCDB, and they were therefore excluded. Our data suggest that population-based CER is less commonly performed in other cancer registries. Second, we identified a nonsignificant trend toward better agreement with the use of instrumental variable and sensitivity analyses; however, only six (2%) of the matched observational studies used propensity score methods, instrumental variable approaches, and sensitivity analyses. Until these methods are routinely used per current recommendations[11,21,22] and a larger sample can be analyzed, it will be difficult to determine whether studies that use all of these methods have significantly better agreement with RCTs. Last, many observational studies did not report median follow-up; therefore, we do not know if differential follow-up in matched pairs influenced agreement.

In summary, we were unable to identify any modifiable factor that improved agreement between population-based observational studies and RCTs beyond chance alone, including statistical rigor, matching populations, or reporting quality. To clinically use population-based CER to infer treatment effects, research is needed to provide quantifiable metrics that improve the reliability and accuracy of this study type.

AFFILIATIONS

[1]Hunter Holmes McGuire VA Medical Center, Richmond, VA
[2]University of Michigan, Ann Arbor, MI
[3]Sohag University Hospital, Sohag, Egypt
[4]University of California San Francisco, San Francisco, CA
[5]Massachusetts General Hospital, Boston, MA

CORRESPONDING AUTHOR

Daniel E. Spratt, MD, University of Michigan Medical Center, Department of Radiation Oncology, 1500 East Medical Center Dr, Ann Arbor, MI 48109-0010; e-mail: sprattda@med.umich.edu.

EQUAL CONTRIBUTION

M.J.S. and D.E.S. contributed equally to this work.

SUPPORT

Supported in part by a Prostate Cancer Foundation Young Investigator Award, Prostate SPORE Grant No. P50CA186786, and generous philanthropic gifts from patients.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST AND DATA AVAILABILITY STATEMENT

Disclosures provided by the authors and data availability statement (if applicable) are available with this article at DOI https://doi.org/10.1200/JCO.18.01074.

AUTHOR CONTRIBUTIONS

Conception and design: Payal D. Soni, Ahmed Abugharib, Steven G. Allen, Matthew J. Schipper, Daniel E. Spratt

Collection and assembly of data: Payal D. Soni, Ahmed Abugharib, Steven G. Allen, Daniel E. Spratt

Data analysis and interpretation: Payal D. Soni, Holly E. Hartman, Robert T. Dess, Steven G. Allen, Felix Y. Feng, Anthony L. Zietman, Reshma Jagsi, Matthew J. Schipper, Daniel E. Spratt

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

REFERENCES

1. Benson K, Hartz AJ: A comparison of observational studies and randomized, controlled trials. N Engl J Med 342:1878-1886, 2000
2. Concato J, Shah N, Horwitz RI: Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 342:1887-1892, 2000
3. Rossouw JE, Anderson GL, Prentice RL, et al: Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. JAMA 288:321-333, 2002
4. Echt DS, Liebson PR, Mitchell LB, et al: Mortality and morbidity in patients receiving encainide, flecainide, or placebo: The Cardiac Arrhythmia Suppression trial. N Engl J Med 324:781-788, 1991
5. Giordano SH, Kuo YF, Duan Z, et al: Limits of observational data in determining outcomes from cancer therapy. Cancer 112:2456-2466, 2008
6. Sacks H, Chalmers TC, Smith H Jr: Randomized versus historical controls for clinical trials. Am J Med 72:233-240, 1982
7. Park HS, Lloyd S, Decker RH, et al: Limitations and biases of the Surveillance, Epidemiology, and End Results database. Curr Probl Cancer 36:216-224, 2012
8. Jagsi R, Abrahamse P, Hawley ST, et al: Underascertainment of radiotherapy receipt in Surveillance, Epidemiology, and End Results registry data. Cancer 118: 333-341, 2012
9. Norris S, Atkins D, Bruening W, et al: Selecting observational studies for comparing medical interventions, in Agency for Healthcare Research and Quality: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD, Agency for Healthcare Research and Quality, 2008
10. Institute of Medicine: Observational Studies in a Learning Health System: Workshop Summary. Washington, DC, National Academies Press, 2013
11. Visvanathan K, Levit LA, Raghavan D, et al: Untapped potential of observational research to inform clinical decision making: American Society of Clinical Oncology research statement. J Clin Oncol 35:1845-1854, 2017
12. Janz TA, Graboyes EM, Nguyen SA, et al: A comparison of the NCDB and SEER database for research involving head and neck cancer. Otolaryngol Head Neck Surg 160:284-294, 2019
13. Lin LI: A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255-268, 1989
14. Ioannidis JP: Why most published research findings are false. PLoS Med 2:e124, 2005
15. Unger JM, Barlow WE, Martin DP, et al: Comparison of survival outcomes among cancer patients treated in and out of clinical trials. J Natl Cancer Inst 106: dju002, 2014
16. Manig L, Käsmann L, Janssen S, et al: Simplified comorbidity score and Eastern Cooperative Oncology Group performance score predicts survival in patients receiving organ-preserving treatment for bladder cancer. Anticancer Res 37:2693-2696, 2017
17. Jackson LA, Nelson JC, Benson P, et al: Functional status is a confounder of the association of influenza vaccine and risk of all cause mortality in seniors. Int J Epidemiol 35:345-352, 2006
18. Boffa DJ, Rosen JE, Mallin K, et al: Using the National Cancer Database for outcomes research: A review. JAMA Oncol 3:1722-1728, 2017
19. McGale P, Cutter D, Darby SC, et al: Can observational data replace randomized trials? J Clin Oncol 34:3355-3357, 2016
20. Bolla M, Van Tienhoven G, Warde P, et al: External irradiation with or without long-term androgen suppression for prostate cancer with high metastatic risk: 10-year results of an EORTC randomised study. Lancet Oncol 11:1066-1073, 2010
21. Jagsi R, Bekelman JE, Chen A, et al: Considerations for observational research using large data sets in radiation oncology. Int J Radiat Oncol Biol Phys 90:11-24, 2014
22. Dreyer NA, Bryant A, Velentgas P: The GRACE checklist: A validated assessment tool for high quality observational studies of comparative effectiveness. J Manag Care Spec Pharm 22:1107-1113, 2016

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Comparison of Population-Based Observational Studies With Randomized Trials in Oncology

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Felix Y. Feng
Leadership: PFS Genomics
Stock and Other Ownership Interests: PFS Genomics
Consulting or Advisory Role: Dendreon, EMD Serono, Janssen Oncology, Ferring Pharmaceuticals, Sanofi
Patents, Royalties, Other Intellectual Property: I helped develop a molecular signature to predict radiation resistance in breast cancer, and this signature was patented by the University of Michigan, my employer; it is in the process of being licensed to PFS Genomics, a company that I helped found (Inst)

Anthony L. Zietman
Leadership: Elsevier

Reshma Jagsi
Employment: University of Michigan
Stock and Other Ownership Interests: Equity Quotient
Consulting or Advisory Role: Amgen, Vizient
Research Funding: AbbVie (Inst)
Travel, Accommodations, Expenses: Amgen

Matthew J. Schipper
Consulting or Advisory Role: Armune Bioscience

Daniel E. Spratt
Consulting or Advisory Role: Janssen, Blue Earth

No other potential conflicts of interest were reported.