Cancer Epidemiology 51 (2017) 49–55

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

A systematic review of instrumental variable analyses using geographic region as an instrument

Emily A. Vertosick[1], Melissa Assel[1], Andrew J. Vickers*

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Avenue, 2nd Floor, New York, NY 10017, United States

ARTICLE INFO

Keywords:
Antineoplastic agents
Geographic region
Instrumental variables
Neoplasms
Observational research
Surgery

ABSTRACT

Background: Instrumental variables analysis is a methodology to mitigate the effects of measured and unmeasured confounding in observational studies of treatment effects. Geographic area is increasingly used as an instrument.

Methods: We conducted a literature review to determine the properties of geographic area in studies of cancer treatments. We identified cancer studies performed in the United States which incorporated instrumental variable analysis with area-wide treatment rate within a geographic region as the instrument. We assessed the degree of treatment variability between geographic regions, assessed balance of measured confounders afforded by geographic area and compared the results of instrumental variable analysis to those of multivariable methods.

Results: Geographic region as an instrument was relatively common, with 22 eligible studies identified, many of which were published in high-impact journals. Treatment rates did not vary greatly by geographic region. Covariates were not balanced by the instrument in the majority of studies. Eight out of eleven studies found statistically significant effects of treatment on multivariable analysis but not for instrumental variables, with the central estimates of the instrumental variables analysis generally being closer to the null.

Conclusions: We recommend caution and an investigation of IV assumptions when considering the use of geographic region as an instrument in observational studies of cancer treatments. The value of geographic region as an instrument should be critically evaluated in other areas of medicine.

1. Introduction

The randomized controlled trial is considered the gold standard for determining comparative treatment effectiveness. However, many treatment comparisons of interest have not been subject to randomized trials, at least in some cases because a trial would be of low feasibility due to expense, or because of low patient tolerance for random assignment. Observational data are often used in place of randomized trials to make inferences about treatment effectiveness, but are prey to confounding bias due to patient selection. There are several statistical methods available to control for measured confounding in observational studies, including multivariable regression and propensity score approaches.

Instrumental variable (IV) methods are an alternative approach that were originally developed by economists and more recently applied for causal inference in healthcare research [1]. The purported advantage of instrumental variable methods over multivariable models or propensity scores is that they balance both observed and unobserved confounders. The method involves identifying an instrument, a variable that is associated with treatment but does not influence outcome, other than through the mechanism of treatment. Randomization is typically an excellent instrument as there is generally a high level of agreement between randomized treatment assignment and treatment received but randomization itself has no direct effect on outcome; observational instrumental variables studies attempt to identify instruments with similar properties to randomization. A classic example of a healthcare instrument is distance from a healthcare facility providing a particular type of intervention. Distance will be a valid instrument where, as is often the case, it affects the likelihood of receiving the intervention but not outcome other than through the effect of treatment on outcome. For instance, a woman living a long way from a mammography facility may be unwilling to travel for a mammogram, but is not otherwise at

Abbreviations: IV, instrumental variables; HRR, hospital referral region; HSA, healthcare service area; RR, relative risk; CI, confidence interval; HR, hazard ratio; PRISMA, preferred reporting items for systematic reviews and meta-analyses; AJCC, American Joint Committee on Cancer; SEER, surveillance, epidemiology, and end results; SES, socioeconomic status; CCI, charlson comorbidity index

* Corresponding author.
E-mail addresses: vertosie@mskcc.org (E.A. Vertosick), asselm@mskcc.org (M. Assel), vickersa@mskcc.org (A.J. Vickers).
[1] Co-first authors.

http://dx.doi.org/10.1016/j.canep.2017.10.005
Received 4 April 2017; Received in revised form 12 September 2017; Accepted 7 October 2017
Available online 14 October 2017
1877-7821/

increased risk of breast cancer death. Ideally, a good instrument satisfies three assumptions: the IV predicts treatment, there is no direct effect of the instrument on outcome except through treatment, and no unmeasured confounding between instrument and outcome [1].

Geographic area has recently been used as an instrument in studies of cancer treatments [2,3]. Through the Dartmouth Atlas program, patterns of use of hospital care known as hospital referral regions (HRR) or healthcare service areas (HSA/HCSA) were established. HSAs are defined by assigning ZIP codes to the hospital area where the greatest proportion of their Medicare residents were hospitalized. HRRs were defined by assigning HSAs to the region where the greatest proportion of major cardiovascular procedures were performed, with minor modifications to achieve geographic contiguity, a minimum population size of 120,000, and a high localization index [4]. Where geographic region is used as the instrument, treatment prevalence within HSAs or HRRs are calculated and used in the estimation of treatment effect [5].

As is well known from the work of the Dartmouth atlas, there are important variations in treatment rates across geographic regions. However, extreme variation across geographic regions would be unexpected. For instance, for the purposes of ensuring that we provide high quality care, it is important to bring to light if 50% of patients in some areas receive a particular treatment compared to 30% of patients in other areas, as this suggests over or undertreatment in at least one of the two areas. But we would be surprised if, say, treatment rates were 80% in one region versus 20% in another.

We hypothesized that geographic region is not strongly associated with cancer treatments and that patient characteristics vary regionally, both of which would suggest that geographic region is not a good instrument. To investigate these hypotheses, we performed a systematic review of studies where geographic region was used as an instrument in an observational study of a cancer treatment. Our aims were to determine the prevalence of this type of analysis, assess the association between treatment and geographic region, document the degree to which geographic instruments balance measured confounders and compare inferences from instrumental variables analyses with traditional multivariable approaches to observational data.

2. Methods

Using PubMed and Google Scholar, we searched for cancer studies, defined as those that examined the effect of at least one treatment, intervention, or screening in cancer patients. The terms “instrumental variable” or “instrumental variables” and “cancer” were searched in conjunction with the following search terms to identify geographic instruments: “hospital service area”, “healthcare service area”, “HCSA”, “HSA”, “hospital referral region”, “HRR”, “geographic area”, and “geographic variation”. Eligible studies had to use instrumental variable analysis with area-wide treatment rate within a geographic region as the instrument, and were limited to those cancer studies using data from the United States. Data collected from these studies included general data on the research question being investigated: indication for treatment or type of cancer; treatment(s) or intervention(s); sample size; definition of geographic area used; geographic area-based instrumental variable; and range of treatment rates between geographic areas. We also collected information on methodology and the types of analyses performed: types of statistical methodology used for instrumental variable and non-instrumental variable analyses; the F-statistic from the test of the association between treatment rate and instrument; outcome(s) of interest; central estimates of effect size with 95% confidence interval for instrumental variable analysis and any other analyses reported, such as propensity score or multivariable analyses; any covariates that remained unbalanced in the instrumental variable analysis.

In many studies, results were presented for more than one outcome, more than one type of statistical methodology, or more than one subset of the cohort. For studies that presented results for multiple outcomes, we chose the outcome for which comparable results on the same scale were available for both instrumental variable and non-instrumental variable analyses. If there was more than one outcome with comparable results, the outcome chosen as primary by the author was selected. If no primary outcome was specified, overall survival was used. If an outcome was assessed at multiple time points, the earliest time point was chosen, as in all studies outcomes at multiple time points were assessed independently, rather than by using longitudinal methods.

For studies which reported multiple IV methodologies we gave preference to the two-stage residual inclusion methodology as 2SRI has been found to provide a less biased result [6,7]; in cases where this was not available we used two-stage least squares methodology. If neither of the two-stage common methodologies was used, we reported the single IV methodology given by the author. In one case IV results were presented using both a binary and categorical categorization of the IV [8]. We reported only the results based on the binary-defined instrument since binary categorization is reported more commonly than categorical subclassifications.

If a study provided more than one type of non-instrumental variable analysis, we chose the methodology that presented estimates of effect size that could be compared to the instrumental variable analysis. Given that standard multivariable models and propensity scores methods tended to give very similar results, we chose to report the multivariable analysis if more than one comparable analysis was performed. If no multivariable models were provided, propensity scores, of whatever form, were used. If a cohort was divided into two or more subsets, the subset with the larger sample size was chosen for the purpose of comparing results.

The reporting of covariate balance was based on specific references by the study authors in the text. If the authors did not mention covariate balance, it was based on p-values ($\alpha = 0.05$) presented in tables showing the balance of covariates across levels of the instrumental variable.

3. Results

The initial search resulted in 279 potentially eligible studies (Fig. 1). Of these studies, 245 were excluded on initial review. Full manuscripts were reviewed for 33 studies, with another 11 studies then excluded. This resulted in 22 eligible studies included in our systematic review. The most commonly studied cancers were prostate cancer, with ten studies, followed by lung and breast cancers. The specific reporting of how instrumental variable analysis was performed varied among the studies. Eleven provided a treatment effect estimate from the instrumental variable analysis and an estimate from a non-instrumental variable analysis that could be compared on the same scale. Fourteen studies reported information on the range of treatment rates seen in each geographic region.

Regardless of how the IV was included in the analyses, treatment rates were most commonly reported as either the mean rate for regions above and below the median value of the instrument or the prevalence in the lowest versus the highest quintile. However, there was considerable variation in reporting. For example, Posner et al. [9] report no summary of treatment rates, Kuo et al. [10] report an overall rate, Hadley et al. [11] and Wright et al. [12] report the average rate above and below the median, and McDowell et al. [13] report the average rate for the highest and lowest quintile. One study did not report treatment rates by low use and high use areas, but reported a range of treatment rates from 0% to 60% over 134 regions, with a median treatment rate of 12% [14]. Most studies report a narrow range of treatment rates, which indicates a weak association between geographic region and treatment. The smallest difference between high and low use areas was 4.8% [8], while the largest difference was 30.7% [15]. The majority of studies reported a difference in treatment rates between high and low use areas of 5% to 20% (9 of 14 studies).

Regardless of the difference in effect size between IV and non-IV analyses, eight out of eleven studies with comparable results rejected

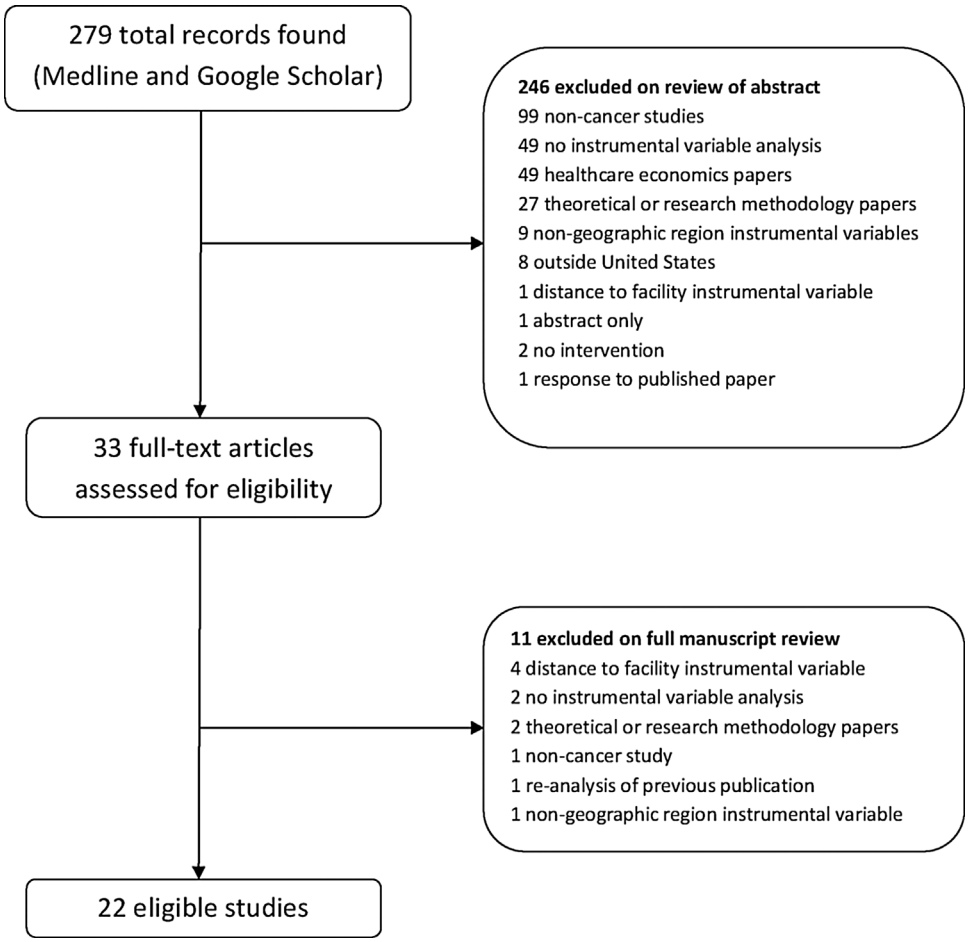

Fig. 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Flow Diagram.

the null hypothesis using non-IV methodology but did not reject the null in the instrumental variable analysis, as described in Table 3. In 9 of 11 studies, the central estimate for IV was closer to the null than the estimate from the multivariable model or propensity score analysis; in 7 studies, the central estimate for IV was closer to the null and outside the 95% C.I. of the multivariable model estimates.

Due to the low variability in treatment rates the standard error in the instrumental variable analysis is expected to be larger than in corresponding non-IV analyses [16,17]. We discovered that a subset of studies had IV standard errors that were not as large as would be expected. When investigating, we found that these studies calculated the standard error using bootstrap methods. The ratio of the IV standard error to the non-IV standard error was 2.0 or less for all studies reporting a bootstrapped standard error. When using a standard IV methodology, the IV standard error was at least 4.8 times greater than the standard error of the corresponding multivariable or propensity score analysis (Table 3). Further research is necessary to explain why the methods used to calculate standard errors resulted in differences of this magnitude.

Fourteen studies reported F-statistics testing the association between treatment and instrument as evidence of the strength of their instruments, with F-statistics ranging from 8.57 to 7792. These high F-statistics are not surprising, given that sample sizes ranged from 1843 to 67,087. Only one study provided any further information on the strength of the instrument, reporting partial r [2] [18].

Sixteen studies provided results that investigated covariate balance across geographic regions, with 13 reporting at least one covariate varying geographically (Table 1). Socioeconomic variables (e.g. race, ethnicity, income, and residence) and disease characteristics/comorbidities were reported as unbalanced across different levels of the instrument in 11 and 9 studies respectively. (Table 1). Six studies failed to report any breakdown of covariate balance by the instrument. Only three studies demonstrate well-balanced covariates by the instrument. In the Kuo et al. study, which reported well-balanced covariates, estimates from IV and non-IV analyses were similar, although the non-IV analysis rejected the null hypothesis while the IV analysis did not, due to a large standard error (IV analysis: relative risk (RR) 1.08 (95% confidence interval (CI) 0.89, 1.31); non-IV analysis: RR 1.11 (95% CI 1.08, 1.13); Table 2) [10]. In the Lu-Yao 2014 et al. study, the IV analysis results were closer to the null and had a larger confidence interval than the non-IV analyses for patients with moderately differentiated disease, despite reporting good covariate balance (IV analysis: hazard ratio (HR) 1.03 (95% CI 0.96, 1.10); non-IV analysis: HR 1.20 (95% CI 1.16, 1.23); Table 2) [19]. The third study reporting good covariate balance did not compare IV to multivariable results [20].

4. Discussion

In our review of the literature, we found that geographic region was commonly used as an instrument in studies of cancer treatment effectiveness. Many of the 22 studies were published in high impact journals, including Journal of the American Medical Association, Journal of the National Cancer Institute and Journal of Clinical Oncology. We found that there was not compelling evidence that geographic region was strongly associated with treatment, with only small differences between high and low use areas. We also found that instrumental variables approaches often did not lead to covariate balance, a violation of one of the assumptions of a good instrument. Furthermore, the estimates of treatment effect were generally closer to the null than for analyses based on multivariable methods. To restate, this is not a general

Table 1
Description of Eligible Studies Included in Analysis, N = 22.

Trial	Indication	Treatment Arms	Sample Size	Instrument	Unbalanced covariates
Basu 2015 [20]	Localized prostate cancer	Radical prostatectomy vs. watchful waiting	11,036	Rate of watchful waiting in the prior year in hospital referral region	Reported good covariate balance
Bekelman 2013 [26]	Muscle-invasive urothelial carcinoma of the bladder	Radical cystectomy vs. bladder-preserving therapy	1843	Rate of radical cystectomy in hospital referral region	Race, area median income, hospital academic affiliation
Bekelman 2015 [27]	Locally advanced prostate cancer	Androgen deprivation therapy vs. androgen deprivation therapy and radiation	12,924	Rate of prostate cancer treatment (surgery or radiation) in hospital referral region	Race, ethnicity, area-level median income
Brooks 2003 [8]	Early stage breast cancer	Breast conserving surgery with radiation vs mastectomy	2905	Local area rate of breast conserving surgery with radiation within 50 mile radius of patient residence	Tumor grade
Brooks 2012 [15]	Early stage breast cancer	Breast conserving surgery with radiation vs mastectomy	28,675	ZIP code specific mastectomy practice styles	Tumor grade, hospital bed size, residence area size, area median income
Earle 2001 [18]	Stage IV non-small cell lung cancer	Chemotherapy within 2 months of diagnosis vs. no chemotherapy	6232	Rate of chemotherapy in health care service area	Race, comorbidities
Gandaglia 2014[28]	Prostate cancer treated with radical prostatectomy	Open radical prostatectomy vs. robotic-assisted radical prostatectomy	5915	Rate of robotic-assisted radical prostatectomy in health service area	IV covariate balance not reported
Hadley 2010 [11]	Early stage prostate cancer	Radical prostatectomy vs. conservative management	14,302	Previous year's (lagged) adjusted probability of receiving conservative management in hospital referral region	Race, marital status, grade, any Medicare claims in previous year (all $p < 0.001$)
Kuo 2012 [10]	Prostate cancer	Primary androgen deprivation therapy vs no therapy	31,930	Proportion of patients receiving primary ADT within each health service area and urologist preference of primary ADT based on previous month's rate	Reported good covariate balance
Lu-Yao 2008 [29]	Localized (T1/T2) prostate cancer	Primary androgen deprivation therapy vs. conservative management	19,271	Rate of primary androgen deprivation therapy in health service area	IV covariate balance not reported
Lu-Yao 2012 [30]	Localized (T1/T2) prostate cancer	Primary androgen deprivation therapy vs. conservative management	29,775	Rate of primary androgen deprivation therapy in health service area	IV covariate balance not reported
Lu-Yao 2014[19]	Localized (T1/T2) prostate cancer	Primary androgen deprivation therapy vs. conservative management	66,717	Rate of primary androgen deprivation therapy in health service area	Reported good covariate balance
McDowell 2014 [13]	Stage I/II pancreatic ductal adenocarcinoma	Pancreatic resection vs. no resection	8323	Rate of resection in health service area	Race, American Joint Committee on Cancer (AJCC) stage, use of radiation ($p < 0.0001$)
Parmar 2013 [14]	Locoregional pancreatic adenocarcinoma	Diagnostic endoscopic ultrasound vs. no endoscopic ultrasound	10,505	Rate of endoscopic ultrasound use in health service area	Race, education, Surveillance, Epidemiology and End Results (SEER) region
Posner 2001 [9]	Early stage breast cancer	Screening mammography vs. no screening	4656	Trichotomous region user status (Atlanta, Seattle, and Connecticut)	IV covariate balance not reported
Saito 2011 [31]	Advanced non-small cell lung cancer	No chemotherapy vs. standard chemotherapy vs. aggressive chemotherapy	7879	Proportion of patients receiving chemotherapy in health care service area	IV covariate balance not reported
Saito 2011 [32]	Advanced non-small cell lung cancer	No hospice vs. short term hospice (< 3 days) vs. long term hospice (4+ days)	7879	Median hospice availability per 1 million population in health care service area	Gender, race/ethnicity, urban/rural residence, year of diagnosis, care in teaching hospital
Sun 2014 [33]	Localized prostate cancer	Radical prostatectomy vs. radiotherapy vs. observation	67,087	Treatment pattern for radical prostatectomy and radiotherapy in health care service area	Age, race, Charlson comorbidity index (CCI), population density, marital status, tumor stage, tumor grade, estimated life expectancy
Sun 2014 [34]	Localized kidney cancer	Radical nephrectomy vs. partial nephrectomy vs. non-surgical management	10,595	Treatment intensity of surgery (Partial and radical nephrectomy) in health care service area	Age, sex, race, CCI, Socioeconomic status (SES), marital status, tumor size, histology, Fuhrman grade
Wisnivesky 2010 [35]	Stage 1 and 2 non-small cell lung cancer	Radiotherapy vs. no treatment	5531	Use of radiotherapy in health care service area	IV covariate balance not reported
Wright 2014 [12]	Advanced-stage ovarian cancer	Neoadjuvant chemotherapy vs. primary surgery	9587	Rate of receipt of primary chemotherapy in hospital referral region	Year of diagnosis, race, SES, histology
Zeliadt 2006 [36]	Non-metastatic prostate cancer	Adjuvant androgen deprivation therapy plus radiotherapy vs. radiotherapy	31,643	Androgen deprivation therapy rates in health care service area	Race, income, comorbidities, age at diagnosis, Gleason at diagnosis, urban area

Table 2
Results From Instrumental Variable and Corresponding Non-instrumental Variable Analyses.

Trial	Treatment Range	Outcome	IV Estimate	Non-IV Estimate	Non-IV Methodology
Basu 2015[20]	Not reported	Adjusted overall survival at 11 years	1.1 months increased adjusted survival with radical prostatectomy (95% CI −25, 28)		
Bekelman 2013[26]	Below median: 73.5%; above median: 81.2%	Adjusted overall survival	1.06 (0.78, 1.31)	1.26 (1.07, 1.50)	Multivariable Cox model
Bekelman 2015[27]	Not reported	Adjusted overall survival	0.65 (0.45, 0.96)	0.63 (0.59, 0.67)	Propensity score
Brooks 2003[8]	Below median: 7.3%; above median: 12.1%	Difference in adjusted overall survival between low and high use areas at 1 year	−0.32	−0.001	Adjusted ordinary least squares
Brooks 2012[15]	Lowest quintile: 37.9%; highest quintile: 68.6%	Adjusted survival rate difference with mastectomy at 7 years	−0.071 (standard error 0.029)	-0.030 (standard error 0.006)	Adjusted survival rate difference with mastectomy at 7 years (linear probability estimates)
Earle 2001[18]	Lowest quintile: 21.2%; highest quintile: 39%	Difference in adjusted overall survival between low and high use areas at 1 year	9% (4%, 23%)		
Gandaglia 2014[28]	Not reported	30-day complications	1.19 (0.97, 1.46)		
Hadley 2010[11]	Below median: 27.1%; above median: 35%	Adjusted overall survival (did not specify overall or cancer-specific survival as primary, cancer-specific survival was HR with error)	1.09 (0.46, 2.59)	1.47 (1.35, 1.59)	Multivariable Cox model
Kuo 2012[10]	Not reported	Relative risk of adjusted overall survival	1.08 (0.89, 1.31)	1.11 (1.08, 1.13)	Multivariable Probit Model
Lu-Yao 2008[29]	Lowest tertile: 30.6%; highest tertile: 52.5%	Adjusted overall survival	1.00 (0.96, 1.05)	1.17 (1.12, 1.21)	Multivariable Cox model
Lu-Yao 2012[30]	Lowest tertile: 31.4%; highest tertile: 47.1%	Adjusted time to use of palliative treatment	1.05 (0.97, 1.14)		
Lu-Yao 2014[19]	Moderately differentiated disease − lowest tertile: 22.3%; highest tertile: 38.8%	Adjusted overall survival	Moderately differentiated: 1.03 (0.96, 1.10)	Moderately differentiated: 1.20 (1.16, 1.23)	Multivariable Cox model
McDowell 2014[13]	Lowest quintile: 38.6%; highest quintile: 67.3%	Adjusted overall survival	0.54 (0.39, 0.75)		
Parmar 2013[14]	Range: 0% to 60%; median 12%	Adjusted overall survival	1.00 (0.73, 1.36)	0.78 (0.73, 0.84)	Multivariable Cox model
Posner 2001[9]	Not reported	Adjusted odds of early stage at diagnosis	3.01 (1.09, 8.34)	2.97 (2.56, 3.45)	Multivariable logistic regression
Saito 2011[31]	Lowest quintile: 30.8%; highest quintile: 52.1%	Difference in adjusted cancer-specific survival between low and high use areas at 1 year	6.7% (−6.6%, 19.9%)		
Saito 2011[32]	Below median: 42.9%; above median: 53.1%	Difference in adjusted overall survival between low and high use areas at 1 year	14.7% (standard error 10.2%)		
Sun 2014 [33]	Not reported	Adjusted overall survival	Radical prostatectomy vs radiation: Life expectancy < 10: 0.81 (0.45, 1.48); Life expectancy > = 10: 0.66 (0.56, 0.79); Radical prostatectomy vs observation: Life expectancy < 10: 0.54 (0.39, 0.75); Life expectancy ≥ 10: 0.59 (0.49, 0.71)		
Sun 2014[34]	Not reported	Adjusted cancer specific mortality	Partial nephrectomy: 0.45 (0.24, 0.83); Radical nephrectomy: 0.58 (0.35, 0.96)		
Wisnivesky 2010[35]	Below median: 50.4%; above median: 64.7%	Difference between the adjusted overall survival in the high and low-use areas divided by the probability of undergoing radiation in those regions at 1 year	15.6% (2.0%−33.9%)		
Wright 2014[12]	Below median: 26.2%; above median: 34.9%	Adjusted overall survival	1.04 (0.67, 1.60)	1.27 (1.19, 1.35)	Multivariable Cox Model
Zeliadt 2006[36]	Below median: 13.4%; above median: 23.4%	Adjusted overall survival	T1/T2: 0.93 (0.82, 1.06)	T1/T2: 1.53 (1.19, 1.90)	Multivariable Cox model

Table 3
Rejection of the null hypothesis by instrumental variable and non-instrumental variable analyses, and ratio of instrumental variable standard error to non-instrumental variable standard error, separately for studies who did and did not use bootstrap correction of instrumental variable standard errors. If a study provided stratified results for both instrumental variable and non-instrumental variable analyses, only one result was included. Studies were included if they provided an estimate of effect size and corresponding standard error for both instrumental variable and non-instrumental variable analyses.

	Studies (Ratio of standard errors)	
Type of IV	IV and non-IV reject null hypothesis	IV does not reject and non-IV rejects null hypothesis
Standard Errors, Bootstrapped (N = 4)	None	Bekelman 2013 (1.2) [26] Lu-Yao 2008 (1.0) [29] Lu-Yao 2014 (2.0) [19] Zeliadt 2006 (0.34) [36]
Standard Errors, Not Bootstrapped (N = 7)	Bekelman 2015 (6.4) [27] Brooks 2012 (4.8) [15] Posner 2001 (8.1) [9]	Hadley 2010 (8.9) [11] Kuo 2012 (8.4) [10] Parmar 2013 (5.7) [14] Wright 2014 (5.8) [12]

critique of IV methods, but an analysis of one particular instrument, geographic region, for one particular application, observational studies on cancer treatments.

One notable feature of the literature is that the number of studies using geographic region as an instrument increased after publication of one high profile paper that expressly recommended this approach as superior to multivariable methods. Hadley et al. reported that the results of an IV analysis comparing surgery to conservative management of prostate cancer (HR 0.73) were more similar to a benchmark randomized trial (HR 0.87) than those of the multivariable approach (HR 1.59). They concluded that "IV analysis may be a useful technique in comparative effectiveness studies of cancer treatments" [11]. However, this result was based on an elementary error, the authors having reversed the reference group in the randomized trial. Hence the results of the multivariable analysis and a propensity score analysis were in fact closer to the benchmark, with those of the IV analyses being in the wrong direction [21]. When correcting for this error, the estimate of the hazard ratio for the IV analysis, when compared to the non-IV analyses, was furthest from the corrected HR for the randomized trial and in the opposite direction. Remarkably, the authors responded that this gross error does not "fundamentally alter our conclusion about the potential strengths and weaknesses of instrumental variable estimation" [22]. Since the publication of this paper, there have been 8 additional publications in the prostate cancer setting using geographic region as an instrument, representing the most common indication (36%) in this systematic review. Several of these papers cited the flawed Hadley study as justification for their approach, presumably because Hadley et al. concluded that the IV result was closest to the result of the randomized controlled trial.

A second paper that explicitly compared the results of an IV analysis to a randomized trial was that of Bekelman et al. [27]. In this case, IV and propensity score estimates were similar to each other and the RCT, except that the IV estimates had a larger standard error.

Several papers justified their choice of instrument by reporting an F-statistic for the association between the instrument and treatment. This approach appears to derive from Staiger and Stock, who stated that "It is our impression that, in applications of two-stage least squares, it is common for the first stage F-statistic … to take on a value less than 10" [23]. We find this reasoning to be invalid, as a high F-statistic might reflect either a strong association in a small study, or a weak association for a large study. This is a particular problem for the sort of observational data sets used by the studies in the current review, which were very large, often greater than 10,000 subjects. F-statistics calculated from such large sample sizes will likely be larger than 10 even if treatment and instrument do not have an important association. The low variability in treatment rates seen across geographic regions in these cancer studies suggests that geography is not a good instrument, despite the high F-statistics reported for individual studies [24].

A high proportion of studies in our review reported that covariates were not balanced across geographic regions. Given that measured and unmeasured confounders are correlated, this finding suggests that geographic region does not balance confounders and therefore that this instrument is not a good surrogate for randomization.

It is possible that our systematic review did not detect some studies using instrumental variables in the cancer setting. It is difficult to search for studies by methodology, and some studies may have referred to an instrument analogous to geographic region using terms that were not included in our search. However, we did identify a large number of studies using geographic region, and found that in no cases did the treatment rate vary greatly by geographic region. The inclusion of only cancer studies in this review may limit the generalizability of our findings. It is of course possible treatment rates may vary geographically much more for treatment of diseases other than cancer, or that covariates for other diseases have better covariate balance, although we see no particular reason why either would be true. We should also note that there are some inherent statistical issues with the use of instrumental variables that go somewhat beyond the scope of this review. In brief, it has been shown that two-stage instrumental variables methods can be biased for binary or time-to-event endpoints [7,25].

The low variability of treatment rates seen in these cancer studies, along with unbalanced characteristics across regions and evidence of unmeasured confounding, demonstrate that treatment rate by geographic region was a problematic instrument. We recommend caution and an investigation of IV assumptions when considering the use of geographic region as an instrument in observational studies of the effects of cancer treatment. Further research should examine the appropriateness of geographic region as an instrument in other healthcare areas.

Authorship contribution

The paper was conceived by Andrew Vickers. Systematic review and assessment of studies was conducted by Emily Vertosick and Melissa Assel. The manuscript was written by Emily Vertosick, Melissa Assel and Andrew Vickers, with all authors approving the final version.

Conflict of interest

None.

Funding

This work was supported by David H. Koch provided through the Prostate Cancer Foundation; the Sidney Kimmel Center for Prostate and Urologic Cancers; SPORE grant from the National Cancer Institute to Dr. H. Scher (grant number P50-CA92629); and a National Institutes of Health/National Cancer Institute Cancer Center Support Grant to MSKCC (grant number P30-CA008748). None of the funding sources had involvement in the conduct of the research or preparation of the manuscript.

References

[1] J.A. Rassen, M.A. Brookhart, R.J. Glynn, et al., Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships, J. Clin. Epidemiol. 62 (12) (2009) 1226–1232.
[2] N.M. Davies, G.D. Smith, F. Windmeijer, et al., Issues in the reporting and conduct

of instrumental variable studies: a systematic review, Epidemiology (Cambridge, Mass) 24 (3) (2013) 363–369.
[3] L.F. Garabedian, P. Chu, S. Toh, et al., Potential bias of instrumental variable analyses for observational comparative effectiveness research, Ann. Intern. Med. 161 (2) (2014) 131–138.
[4] J. Wennberg, Appendix on the geography of health care in the United States, Dartmouth Atlas Health Care U. S. 28 (2015) 9–96.
[5] O. Klungel, J. Uddin, A. de Boer, et al., Instrumental variable analysis in epidemiologic studies: an overview of the estimation methods, Pharmaceutica Analytica Acta 6 (2015) 353.
[6] J.V. Terza, A. Basu, P.J. Rathouz, Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling, J. Health Econ. 27 (3) (2008) 531–543.
[7] B. Cai, D.S. Small, T.R. Have, Two-stage instrumental variable methods for estimating the causal odds ratio: analysis of bias, Stat. Med. 30 (15) (2011) 1809–1824.
[8] J.M. Brooks, E.A. Chrischilles, S.D. Scott, et al., Was breast conserving surgery underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa, Health Serv. Res. 6 (Pt 1) (2003) 1385–1402.
[9] M.A. Posner, A.S. Ash, K.M. Freund, et al., Comparing standard regression, propensity score matching, and instrumental variables methods for determining the influence of mammography on stage of diagnosis, Health Serv. Outcomes Res. Methodol. 2 (3-4) (2001) 279–290.
[10] Y.F. Kuo, J.E. Montie, V.B. Shahinian, Reducing bias in the assessment of treatment effectiveness: androgen deprivation therapy for prostate cancer, Med. Care 50 (5) (2012) 374–380.
[11] J. Hadley, K.R. Yabroff, M.J. Barrett, et al., Comparative effectiveness of prostate cancer treatments: evaluating statistical adjustments for confounding in observational data, J. Natl. Cancer Inst. 102 (23) (2010) 1780–1793.
[12] J.D. Wright, C.V. Ananth, J. Tsui, et al., Comparative effectiveness of upfront treatment strategies in elderly women with ovarian cancer, Cancer 120 (8) (2014) 1246–1254.
[13] B.D. McDowell, C.G. Chapman, B.J. Smith, et al., Pancreatectomy predicts improved survival for pancreatic adenocarcinoma: results of an instrumental variable analysis, Ann. Surg. 261 (4) (2015) 740–745.
[14] A.D. Parmar, K.M. Sheffield, Y. Han, et al., Evaluating comparative effectiveness with observational data: endoscopic ultrasound and survival in pancreatic cancer, Cancer 119 (21) (2013) 3861–3869.
[15] J.M. Brooks, E.A. Chrischilles, M.B. Landrum, et al., Survival implications associated with variation in mastectomy rates for early-staged breast cancer, Int. J. Sur. Oncol. (2012) 2012.
[16] M. Baiocchi, J. Cheng, D.S. Small, Instrumental variable methods for causal inference, Stat. Med. 33 (13) (2014) 2297–2340.
[17] J. Bound, D.A. Jaeger, R.M. Baker, Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak, J. Am. Stat. Assoc. 90 (430) (1995) 443–450.
[18] C.C. Earle, J.S. Tsai, R.D. Gelber, et al., Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis, J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol. 19 (4) (2001) 1064–1070.
[19] G.L. Lu-Yao, P.C. Albertsen, D.F. Moore, et al., Fifteen-year survival outcomes following primary androgen-deprivation therapy for localized prostate cancer, JAMA Int. Med. 174 (9) (2014) 1460–1467.
[20] A. Basu, J.L. Gore, Are elderly patients with clinically localized prostate cancer overtreated? exploring heterogeneity in survival effects, Med. Care 53 (1) (2015) 79–86.
[21] A.J. Vickers, Re Comparative effectiveness of prostate cancer treatments: evaluating statistical adjustments for confounding in observational data, J. Natl. Cancer Inst. 14 (2011) 1134 (author reply −5).
[22] J. Hadley, M.J. Barrett, D.F. Penson, et al., Response: Re: comparative effectiveness of prostate cancer treatments: evaluating statistical adjustments for confounding in observational data, J. Natl. Cancer Inst. 103 (14) (2011) 1134–1135.
[23] D.O. Staiger, J.H. Stock, Instrumental variables regression with weak instruments, Econometrica 65 (3) (1997) 557–586.
[24] D.A. Lawlor, K. Tilling, Davey smith G: triangulation in aetiological epidemiology, Int. J. Epidemiol. 45 (6) (2016) 1866–1886.
[25] F. Wan, D. Small, J.E. Bekelman, et al., Bias in estimating the causal hazard ratio when using two-stage instrumental variable methods, Stat. Med. 34 (14) (2015) 2235–2265.
[26] J.E. Bekelman, E.A. Handorf, T. Guzzo, et al., Radical cystectomy versus bladder-preserving therapy for muscle-invasive urothelial carcinoma: examining confounding and misclassification biasin cancer observational comparative effectiveness research, Value Health: J. Int. Soc. Pharmacoeconom. Outcomes Res. 16 (4) (2013) 610–618.
[27] J.E. Bekelman, N. Mitra, E.A. Handorf, et al., Effectiveness of androgen-deprivation therapy and radiotherapy for older men with locally advanced prostate cancer, J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol. 33 (7) (2015) 716–722.
[28] G. Gandaglia, J.D. Sammon, S.L. Chang, et al., Comparative effectiveness of robot-assisted and open radical prostatectomy in the postdissemination era, J. Clin. Oncol.: Off. J. Am. Soc. Clin. Oncol. 32 (14) (2014) 1419–1426.
[29] G.L. Lu-Yao, P.C. Albertsen, D.F. Moore, et al., Survival following primary androgen deprivation therapy among men with localized prostate cancer, JAMA 300 (2) (2008) 173–181.
[30] G.L. Lu-Yao, P.C. Albertsen, H. Li, et al., Does primary androgen-deprivation therapy delay the receipt of secondary cancer therapy for localized prostate cancer? Eur. Urol. 62 (6) (2012) 966–972.
[31] A.M. Saito, M.B. Landrum, B.A. Neville, et al., The effect on survival of continuing chemotherapy to near death, BMC Palliative Care 10 (2011) 14.
[32] A.M. Saito, M.B. Landrum, B.A. Neville, et al., Hospice care and survival among elderly patients with lung cancer, J. Palliat. Med. 8 (2011) 929–939.
[33] M. Sun, J.D. Sammon, A. Becker, et al., Radical prostatectomy vs radiotherapy vs observation among older patients with clinically localized prostate cancer: a comparative effectiveness evaluation, BJU Int. 2 (2014) 200–208.
[34] M. Sun, A. Becker, Z. Tian, et al., Management of localized kidney cancer: calculating cancer-specific mortality and competing risks of death for surgery and non-surgical management, Eur. Urol. 65 (1) (2014) 235–241.
[35] J.P. Wisnivesky, E. Halm, M. Bonomi, et al., Effectiveness of radiation therapy for elderly patients with unresected stage I and II non-small cell lung cancer, Am. J. Respir. Crit. Care Med. 3 (2010) 264–269.
[36] S.B. Zeliadt, A.L. Potosky, D.F. Penson, et al., Survival benefit associated with adjuvant androgen deprivation therapy combined with radiotherapy for high- and low-risk patients with nonmetastatic prostate cancer, Int. J. Radiat. Oncol. Biol. Phys. 2 (2006) 395–402.