Published in final edited form as:
Stat Biosci. 2011 September ; 3(1): 6–27. doi:10.1007/s12561-011-9033-6.

Estimating Decision-Relevant Comparative Effects Using Instrumental Variables

Anirban Basu
Departments of Health Services and Pharmacy, University of Washington, Seattle, 1959 NE Pacific St, Box 357660, Seattle, WA 98195-7660, USA

Anirban Basu: basua@uw.edu

Abstract

Instrumental variables methods (IV) are widely used in the health economics literature to adjust for hidden selection biases in observational studies when estimating treatment effects. Less attention has been paid in the applied literature to the proper use of IVs if treatment effects are heterogeneous across subjects. Such a heterogeneity in effects becomes an issue for IV estimators when individuals' self-selected choices of treatments are correlated with expected idiosyncratic gains or losses from treatments. We present an overview of the challenges that arise with IV estimators in the presence of effect heterogeneity and self-selection and compare conventional IV analysis with alternative approaches that use IVs to directly address these challenges. Using a Medicare sample of clinically localized breast cancer patients, we study the impact of breast-conserving surgery and radiation with mastectomy on 3-year survival rates. Our results reveal the traditional IV results may have masked important heterogeneity in treatment effects. In the context of these results, we discuss the advantages and limitations of conventional and alternative IV methods in estimating mean treatment-effect parameters, the role of heterogeneity in comparative effectiveness research and the implications for diffusion of technology.

Keywords

Instrumental variables; Local IV methods; Heterogeneity; Breast cancer; Survival

1 Introduction

Recent legislation around investments in comparative effectiveness research (CER) has raised awareness and enthusiasm for the development of methods for such research. A contemporaneous investment in health information technology has raised hopes for the development of richer and comprehensive observational databases based on electronic medical records. Despite the push for the larger use of such databases in CER, the fundamental methodological challenge of selection bias arising out of non-random assignment of treatments remains. Since the goal of CER is to generate information that can inform better treatment selection in practice, causal estimation of treatment effects remain central to the CER theme. Otherwise, interventions that do not provide sufficient value may be adopted and treatments that do may be eliminated.

Selection bias (i.e., confounding by indication) arises when factors that can influence the treatment choice such as patient health and provider skills also influence outcomes. This is a

Correspondence to: Anirban Basu, basua@uw.edu.

common phenomenon in observational studies of treatment outcomes. The significance of this well-known limitation was famously illustrated in the case of hormone replacement therapy in post-menopausal women. As several large scale observational studies consistently showed these treatments to be effective for preventing chronic cardiovascular disease, hormone replacement therapy became widely adopted. Use then plummeted when these studies were eventually disproven by a large randomized trial [35]. It has subsequently been shown that the reason for the discrepant results was that the observational studies failed to consider certain confounders like socioeconomic status [25] or failed to distinguish initiation of therapy from prevalence of therapy [24]. The significance of overcoming the limitation of common observational study designs cannot be overstated as it could lead to fewer mistaken conclusions regarding treatment effectiveness and a greater use of sound observational studies to develop the evidence base of comparative effectiveness research.

A wide range of statistical methods have been developed to address overt selection bias or bias that arise due to differences in levels of confounders for patients receiving different treatments that are observed by the analyst of the observational data. Some of the most common techniques used to address overt bias include regression methods, propensity score matching and doubly robust estimators [3, 34, 36, 38, 40]. The set of techniques that rely on propensity scores and related techniques that ensure balance of confounders between groups are being widely adopted in comparative effectiveness research as they often provide better estimates of treatment effects [39] and can be implemented across a wide range of settings using data readily available. However, these methods have limitations if confounders that are not observed by the analysts give rise to hidden selection bias [43, 44]. This hidden selection bias presents the biggest challenge for comparative effectiveness research as aptly illustrated in the hormone replacement therapy example.

Because of the prevalence of hidden selection bias, instrumental variable (IV) analysis has been a cornerstone method for observational studies, whose origins date back to the 1920s [42]. In the last couple of decades, these methods have gained popularity in the medical literature on the evaluation alternative medical treatments [9, 10, 14, 27, 43], the types of evaluations that were by and large restricted to clinical trials. The instrumental variables determine or affect treatment choice, but do not have a direct effect on outcomes except to the extent that they influence the choice of treatment [1, 2, 16]. Thus, by using IVs, one can induce substantial variation in the treatment variable but have no direct effect on the outcome variable of interest. One can then estimate how much of the variation in the treatment variable is induced by the instrument—and only that induced variation—affects the outcome measure. In econometric terminology, this induced variation is called the *exogenous variation* and identifies the desired estimate. These analyses constitute an important body of work that have advanced the field of CER by going beyond establishing associations between treatments and outcomes to estimating causal effects of treatments on outcomes, such as a RCT conducted on a similar population can inform. The adoption of these techniques for CER, although limited thus far, appears to be accelerating.

The field of CER itself is also grappling with issues about heterogeneity of treatment effects. In many situations, people respond differently to the same treatment. This is called *response heterogeneity*. More importantly, the differential response from alternative treatments may vary across people. This is called *treatment-effect heterogeneity*, and will be the primary focus of discussion in this paper. There are strong economic reasons why heterogeneity is important is this field [4, 6]. But what has received less attention is how such a heterogeneity can compromise the traditional evidence generation infrastructure (e.g. randomized clinical trials and observational data analyses) in CER.

Let us take the case of IV approaches. An IV estimate of treatment effect using standard methods (e.g. two-stage least squares) is comparable to that arising from an RCT only under the assumption that treatment effects are constant for everyone in the population with the same observed characteristics. Even if treatment effects are allowed to be heterogeneous, IV estimates assume patients or their physicians do not have any additional information beyond what the analyst of an observational data possesses that can enable them to anticipate these effects and to select a treatment that would potentially give them the largest benefits. Such assumptions are clearly a stretch for modeling treatment choices in health care, especially under the practical limitations of observational data to collect all relevant information pertaining to treatment choices. Note that such assumption are also implicitly made in RCTs where selection into RCTs are hardly ever studied, even though there are several instances where clinicians have questioned the generalizability of RCTs [12].

When such assumptions are relaxed, recent econometric literature has demonstrated several limitations of the traditional and newer IV approaches that we discussed above [16, 17]. Now subjects and their providers are able to self-select treatments based on the patient's expected idiosyncratic gains, i.e. it allows unobserved characteristics of patients that influence treatment choices to also be moderators of treatment effects (I will later develop a weaker assumption than self-selection that can also lead to such moderation). Imbens and Angrist [26] showed that standard IV methods can identify parameters that reflect the treatment effects for a group of marginal patients, i.e. the patients whose actual treatment choices are driven by the specific instrumental variables, but are otherwise indifferent to choosing between alternative treatments. Therefore, the marginal patients identified by an IV are entirely dependent on the specific instrument being used and how this instrument affects treatment choices [2, 16]. Consequently, the use of different instruments will produce different treatment effects because they represent the effects for different groups of marginal patients, and IV results become instrument dependent. This key insight, originally highlighted by Heckman [15], is that it is difficult to interpret and apply IV results to clinical practice, where patients are often believed to select treatment based on their idiosyncratic net gains or preferences. In response to this insight, most traditional IV methods estimate a Local Average Treatment Effect (LATE). This estimate is often substantially different from mean treatment-effect concepts such as the Average Treatment Effect (ATE). This result, in one sense, is synonymous to the problems of interpreting RCT results, when self-selection into RCTs is common. In fact, under heterogeneity and self selection, even if results from IV methods applied to observational data and results from an RCT are both internally valid, there is no reason to expect that these results should tally with each other. Yet much of the applied literature has tried to replicate RCT results with IV methods.

To recover the full distribution of treatments effects across all possible margins of patients choices, not just the one directly influenced by an IV, one needs to explicitly develop a choice model for treatment selection. This choice model tries to explain choices based on all observed risk factors and also all possible IVs that are identified in the data, so that for each predicted level of probability for treatment choice, we observe some patients choosing treatment and some that do not. One can then study how the difference in average outcomes, the marginal treatment effect (MTE), between these two groups varies over levels of the probability of treatment choice. This approach, known as the local instrumental variable (LIV) approach, uses control function methods to identify the MTEs and subsequently combines them to form interpretable and decision-relevant parameters of interest such as the ATE or the Effect on the Treated (TT) or the Untreated (TUT). (Heckman series) ATE estimates the average gain if everyone undergoes treatment as compared to an alternative treatment or no treatment at all. This has been one of the most popular parameters of interest for health economists and policy analysts when making inference about health care policies [46]. Treatment Effect on the Treated (TT) estimates the average gain to those who actually

select into treatment and is one ingredient for determining whether a given treatment should be shut down or retained as a medical practice or in the formularies. It is informative on the question of whether the persons, choosing the treatment, benefit from it in gross terms. Recently, Basu et al. [5] applied these methods to estimate ATE and TT of breast cancer treatments on costs.

In this paper, my goal would be to highlight these challenges in the context of using instrumental variable methods on observational data and discuss potential solutions to these problems.

2 A Motivating Example

Several RCTs compared survival rates for breast-conserving surgery with radiation therapy (BCSRT) and mastectomy (MST) in the treatment of women with localized (stage 1 or 2) breast cancer. The largest trial (1843 women of all ages evenly divided among the treatment arms) found that there were no statistically significant differences in five-year survival rates, which were 75.9% for MST, and 79.8% for BCSRT [11]. Other RCTs, which had smaller enrollments and compared survival over 6–15 years between BCSRT and MST, also found statistically insignificant survival differences ranging from −8% to 3% [41]. In 2003, Hadley and colleagues used data for a sample of Medicare beneficiaries (age 67 and older) who were treated for localized breast cancer between 1992 and 1994 and analyzed 3-year survival rates using a set of valid IVs and a traditional IV approach [14]. They found that the IV approach produced a comparative effect estimate of −5 percentage points favoring MST that did not reach statistical significance (std. err. = 0.10). The authors discuss several issues regarding the generalizability of the RCT results and its comparison to their IV estimates. More importantly, they point out that IV estimators, though inefficient in the moderate sample size that they utilize, do produce results that conform to contemporaneous RCT evidence. The point estimates of the comparative effect in this scenario arising out of the CER studies and the statistical insignificance of those estimates seem to have conveyed a sense of equivalence for both treatments for all patients. Consequently, the proportion of patients undergoing breast-conserving surgery increased from 41% in 1992 to 60% in 2003, whereas the mastectomy rate decreased from 59% in 1992 to 40% in 2003 ($P < 0.0001$) [47].

Later in this paper, we are going to use the same data as used by Hadley et al. to closely reproduce their results but also establish the distribution of treatment effects in the population and discuss whether such large scale uptake of BSCRT could have been beneficial.

2.1 A Model for Potential Outcomes and Selection

We start by formally developing structural models of outcomes and treatment choice. For the sake of simplicity we will restrict our discussion to two treatment states—the *treated* state denoted by $j = 1$ and the *untreated* state denoted by $j = 0$, and their corresponding potential outcomes represented by

$$Y_1=\mu_1(X,W)+U_1 \tag{1a}$$

$$Y_0=\mu_0(X,W)+U_0 \tag{1b}$$

where $\mu_j(X, W)$ is an unknown nonlinear function of observable (X) and unobservable (W) characteristics and U_j are purely random errors. The fact that Y_0 (or Y_1) vary by levels of X and W indicate that absolute response to a treatment is heterogeneous (i.e. there is response heterogeneity). Conditional on specific levels of X and W, idiosyncratic gains (or losses) from treatment over control is given by $\mu_1(x, w) - \mu_0(x, w)$. These idiosyncratic gains or losses may vary either over observed characteristics X or over unobserved characteristics W or both, giving rise to treatment-effect heterogeneity. The terms *observable* and *unobservable* pertain to the analyst's perspective and these covariates enter the structural model symmetrically in determining potential outcomes [30]. We will refer to this formulation of the symmetric structural nonlinear model as the *pure* nonlinear model. Following standard assumptions in the potential outcomes literature, we posit that $X, W \perp\!\!\!\perp U_j$ and $X \perp\!\!\!\perp W$ where $\perp\!\!\!\perp$ implies statistical independence.

Let D be an indicator that takes the value 1 if an individual selects into treatment and 0 if she does not. Treatment selection is assumed to be driven by levels of both X and W, making them confounders. However, no formal model for treatment choice is required at this point. Each subject either receives treatment or not and the observed outcome becomes $Y = DY_1 + (1 - D)Y_0$. This representation is Quandt's switching regression framework [31, 32]. Consequently, the model for potential outcomes in (1a) and (1b) can be used to obtain a model for the observed outcome:

$$\begin{aligned} Y &= \mu_0(x, w) + D(\mu_1(x, w) - \mu_0(x, w)) + \{D(U_1 - U_0) + U_0\} \\ &= g(D, X, W) + \{D(U_1 - U_0) + U_0\} \end{aligned} \tag{2}$$

Since U_j are purely random errors, the error term $\{D(U_1 - U_0) + U_0\}$ is also purely random and is not the source of hidden biases. This is in sharp contrast to linear models where the unobserved factors generating hidden biases reside in this additive error term. Instead, in a pure nonlinear model, the endogeneity arise due to unobserved factors W that resides within the mean function of the outcome, symmetrically as other observed factors. Since $W \not\perp\!\!\!\perp D$ (i.e., levels of W are different among those who select into treatment versus those who do not) it is not possible to decompose $g(.)$ into additively separate part comprising of the observed and unobserved components unless $\mu_0(.)$ and $\mu_1(.)$ follow an additively separable specification in X's and W's and therefore lend themselves to be used as an ordinary least squared estimator.

The ATE conditional on $X = x$, is given by

$$E(\Delta|X=x) = \int (\mu_1(x, w) - \mu_0(x, w)) dF(w) \tag{3}$$

It estimates the average gain if everyone with characteristics $X = x$ undergoes treatment as compared to remaining untreated [8], and informs whether, on average, a new treatment should replace an older treatment or a no treatment policy. Similarly, another useful parameter that has significant policy relevance in health care is the effect of Treatment on the Treated (TT) which informs whether the person choosing the treatment benefits from it. TT conditional on $X = x$ is formally defined by

$$\text{TT}(x) = E(\Delta|X=x, D=1) = \int (\mu_1(x, w) - \mu_0(x, w)) dF(w|D=1) \tag{4}$$

Notice that, conditional on $X = x$, TT is different from ATE only when the individual-level treatment-effects vary over unobserved confounders W. This kind of (treatment-effect) heterogeneity is termed as "essential" [16]. Essential heterogeneity can arise in two ways. (1) Subjects anticipate this heterogeneity and select treatment based on it. This is the economic self-selection behavior that Heckman and colleagues discuss [16, 19, 23]. (2) Subjects cannot anticipate idiosyncratic gains but select treatment based on W, which determines response heterogeneity $\mu_0(x, W)$ or $\mu_1(x, W)$. However, treatment-effect heterogeneity (or idiosyncratic gains, $\mu_1(x, W) - \mu_0(x, W)$) is NOT independent of response heterogeneity. Thus, even though subjects do not self select based on idiosyncratic gains, their choices and idiosyncratic gains are no longer independent.

When either of these two situations is not met and the distribution of *idiosyncratic gains* ($\mu_1(x, w) - \mu_0(x, w)$) is independent of D, then it leads to the case where treatment effects are heterogeneous but not necessarily essential, making TT(x) and ATE(x) identical.[1]

2.2 Estimation Using Instrumental Variables

Instrumental variables can be used to salvage certain treatment effects. IV analysis tries to model the dependence of unobserved characteristics that influence both treatment choice and outcomes using factors (Z) that influence treatment choice but are not contained in X. In this pure nonlinear model, the treatment effects are always heterogeneous over unobservables with a strong possibility that subjects' choices may be dependent on idiosyncratic gains resulting in essential heterogeneity. An IV assumes

$$Z \perp\!\!\!\perp W \Rightarrow Z \perp\!\!\!\perp (\mu_1(x, W) - \mu_0(x, W)) | X = x \quad \text{(Assumption 1)} \tag{5}$$

The conditional instrumental variable effect, $E(\Delta_{IV}|X)$, for any two values of an instrument, z and z', is given by[2]

$$\lim_{z \to z'} E(\Delta_{IV}|X) = \frac{\partial E(Y|X, Z=z)}{\partial Z} \Big/ \frac{\partial \Pr(D=1|X, Z=z)}{\partial Z} \tag{6}$$

The parameter is the ratio of the change in the conditional expectations of outcomes with respect to Z to the change in the probability of receiving treatment with respect to Z.

2.2.1 Under Non-essential Heterogeneity—If treatment effects are non-essential, then the IV effect estimator in (6) is consistent for the average treatment effect $E(\Delta|X = x)$. This is because, under non-essential heterogeneity, and following assumption (1),

$$D \perp\!\!\!\perp (\mu_1(x, w) - \mu_0(x, w)) \Rightarrow Z \perp\!\!\!\perp D \cdot \{\mu_1(x, w) - \mu_0(x, w)\} \quad \text{(Assumption 2)} \tag{7}$$

Consequently, for any two values on an instrument, z and z', the IV estimator is given by

[1]Under essential heterogeneity, ATE(x) ≠ TT(x), but ATE may be equal to TT, while under non-essential heterogeneity ATE(x) = TT(x) but ATE may not be equal to TT. The unconditional effects depends on $F(X)$ and $F(X|D)$.

[2]For a continuous instrumental variable, the overall IV effect is a weighted average of all possible pairs of values for that instrument. This is further explained below.

$$E(\Delta_{IV}|X=x)=\frac{E(Y|X=x,Z=z')-E(Y|X=x,Z=z)}{\Pr(D=1|X=x,Z=z')-\Pr(D=1|X=x,Z=z)} \tag{8}$$

Based on (2),

$$\begin{aligned}E(Y|X=x,Z=z)&=\int\mu_0(x,w)dF(w|z)+\Pr(D=1|X=x,Z=z)\times\int(\mu_1(x,w)-\mu_0(x,w))dF(w|z,D=1)\\&=\int\mu_0(x,w)dF(w)+\Pr(D=1|X=x,Z=z)\times\int(\mu_1(x,w)-\mu_0(x,w)dF(w),\end{aligned}$$

where the last equality is due to assumptions (1) and (2). Therefore,

$$\begin{aligned}&E(Y|X=x,Z=z')-E(Y|X=x,Z=z)\\&=\{\Pr(D=1|X=x,Z=z')-\Pr(D=1|X=x.Z=z)\}\times\int(\mu_1(x,w)-\mu_0(x,w))dF(w),\quad\text{and}\\&E(\Delta_{IV}|X=x)=\int(\mu_1(x,w)-\mu_0(x,w))dF(w)=E(\Delta|X=x)\end{aligned} \tag{9}$$

Note that in this traditional application of an IV estimator, one can be agnostic about a formal choice model. This, in principle, represents an attractive feature. However, as we show below, absence of an explicit choice model is also a drawback of the IV approach, when the non-essential treatment-effect heterogeneity assumption is relaxed.

2.2.2 Under Essential Heterogeneity—When treatment-effect heterogeneity is essential, the IV estimator in (6) produces a local average treatment effect (LATE), which is the average treatment effect for individuals who would change their treatment choice when Z moves from z to z'. This is because, even if assumption (1) is met under essential heterogeneity, assumption (2) is not. That is, conditionally receiving treatment the IVs may no longer be independent of the idiosyncratic gains in that subgroup ($E(Y_1 - Y_0|D, x, z') \neq E(Y_1 - Y_0|D, x, z)$). Consequently, the inferences based on traditional IV methods breaks down as the effect identified by IV now depends on the practically unidentified margin of patients among whom the change in IV levels can hypothetically induce change in treatment choice. However, the subpopulation induced to change treatment due to changes in levels of instrument is not clearly identified since, in the absence of an underlying choice model, the relevant margin at which this change in behavior is taking place is not specified. Unfortunately, targeting clinical practice or policy to this margin of patients is often difficult, if not impossible, due to lack of explicit identity for these patients.

In order to understand what LATE estimates and how one can go beyond LATE to recover decision-centered parameters such as ATE and TT, one must formulate a formal model for treatment choices that can formally identify the margin of patients influenced by an IV.

2.3 Formal Model for Treatment Choices and Its Link to IV Estimators

2.3.1 The Random Utility Framework—Let the net (latent) utility for treatment,[3] A, based on which choices are determined,[4] be given as

[3]Latent utility in this framework is an anticipated form of utility rather than an experienced form and implicitly accounts for decision maker's preferences which varies over all factors. A factor cannot affect treatment choice unless it affects this latent utility.
[4]Decision maker in a clinical context may as well be the physician–patient dyad and not the patient or the physician alone.

$$\Lambda=\mu_{\Lambda'}(X,W,Z)+U_{\Lambda'}, \tag{10}$$

where, similar to the potential outcomes model in (1), $\mu_{\Lambda'}(X, W, Z)$ is an unknown nonlinear function of observable (X, Z) and unobservable (W) characteristics and $U_{\Lambda'}$ are random errors. Under assumptions of exogeneity for Z and X, which implies that $X, Z \perp\!\!\!\perp U_{V'}$ and $X, Z \perp\!\!\!\perp W$, we can rewrite (10) as

$$\Lambda=\mu_{\Lambda}(X,Z)+U_{\Lambda}, \quad E(U_{\Lambda})=0, \quad D=I(\Lambda>0), \tag{11}$$

where $\mu_{\Lambda}(X, Z) = \int \mu_{\Lambda'}(x, w, z)\, dF(w)$ and $U_{\Lambda} = \Lambda - \mu_{\Lambda}(X, Z)$ has expectation of zero while $I(.)$ is an indicator function representing treatment choice D. Equation (11) expresses the typical random utility framework for discrete choices in econometrics [28, 29]. Following this framework, one can write

$$\begin{aligned} D=I(\Lambda>0)=1(U_{\Lambda}>-\mu_{\Lambda}(z,x)) &\iff 1(F_{U_{\Lambda}}(U_{\Lambda})>F_{U_{\Lambda}}(-\mu_{\Lambda}(z,x))) \\ &\iff 1(F_{U_{\Lambda}}(U_{\Lambda})>1-P(z,x)) \end{aligned}$$

where $P(z, x) = F_{U_{\Lambda}}(\mu_{\Lambda}(z, x))$ and $F_{U_{\Lambda}}(U_{\Lambda}) = U_D \sim \text{Uniform}(0, 1)$ by construction. The formulation in (11) decomposes factors that determine choice of treatment into the observed and unobserved components (again, by the analyst). The additive separability of (10) in terms of observables and unobservables plays a crucial role in the justification of instrumental variable methods [19, 23]. Hereon, we denote $S(z, x) = 1 - P(z, x)$. Consider for simplicity the single instrument case, i.e. Z is a scalar rather than a vector of instruments. Given model (11) and the assumed independence of Z and U_V, changing Z externally from U_V, shifts all people in the same direction (toward or against $D = 1$). This produces "monotonicity" in the sense of Imbens and Angrist [26].

2.3.2 Interpretation of IV Estimators—Armed with a choice model, one can then start to understand the heterogeneity in treatment effects across different margins of the patient population. Recall that U_D represents the unobserved characteristics that determine treatment. Once we condition on the observed factors X and the unobserved U_D, the conditional mean treatment effects $E(\Delta \mid X = x, U_D = u_D, Z = z)$ are exactly the same for each individual with the same value of $U_D = u_D$, despite having different values of Z (or $P(Z, X)$). For any value of the instrument $Z = z$ (and $X = x$), the patients for whom $U_D > S(z, x)$ receive treatment while patients with $U_D \leq S(z, x)$ remain untreated. In addition, notice that the expected value of the observed outcome for this group of patients can be written as the weighted average of those who receive treatment and those who do not:

$$\begin{aligned} E(Y|Z=z,X=x)&=\Pr(D=1|Z=z,X=x)\cdot E(Y_1|D=1,Z=z,X=x)+\Pr(D=0|Z=z,X=x)\cdot E(Y_0|D=0,Z=z,X=x) \\ &=\Pr(U_D>S(z,x))\cdot E(Y_1|U_D>S(z,x),x)+\Pr(U_D\leq S(z,x))\cdot E(Y_0|U_D\leq S(z,x),x) \end{aligned} \tag{12}$$

By the definition of an instrument (Assumption 1), we can vary the value of $Z = z$ (given $X = x$), and therefore $P(z, x)$ and $S(z, x)$, non-trivially with respect to the distribution of U_D. Thus, consider two groups of patients, one with $Z = z$ and the other with $Z = z'$ from the same distribution of U_D. Let $S(z, x) \geq S(z', x)$ for every patient. Using expression (12), we see that the difference in the observed outcomes between these two groups of patients is then

$$E(Y|z,x) - E(Y|z',x) = [\Pr(U_D > S(z,x)) \cdot E(Y_1|U_D > S(z,x),x) + \Pr(U_D \le S(z,x)) \cdot E(Y_0|U_D \le S(z,x),x)] - [\Pr(U_D > S(z',x)) \cdot E(Y_1|U_D > S$$
$$= \Pr(S(z',x)) < U_D < S(z,x)) \times [E(Y_1|S(z',x) < U_D < S(z,x),x) - E(Y_0|S(z',x) < U_D < S(z,x)$$
$$= \Pr(P(z,x) < U_D < P(z',x)) \times [E(Y_1|S(z',x) < U_D < S(z,x),x) - E(Y_0|S(z',x) < U_D < S(z,x),$$

(13)

where the last two equalities follow from the fact that $U_D \sim$ Uniform(0, 1). The mean potential outcomes outside the limits of the margin, $S(z', x) < U_D < S(z, x)$, cancel out. Combining (8) and (13), we can conclude that LATE identifies the average effect for a group of patients who are within the margin defined by $S(z, x)$ and $S(z', x)$ [19, 22]:

$$\text{LATE}(x,z,z') = E(Y_1 - Y_0|X=x, S(z',x) < U_D < S(z,x)) \quad (14)$$

LATE is often referred in the health literature as the treatment effect for the marginal patients [9, 27]. The marginal patients are defined as the subset of patients whose treatment choices varies with the instrument. Imbens and Angrist [26] define the LATE parameter from hypothetical manipulation of the choice probability or values for the instrument. Heckman and Vytlacil [19, 22] draw on choice theory and derive LATE (and also other treatment-effect parameters, as explained below) in the context of the generalized Roy Model [18, 37]. Relating IV to choice models helps to identify the margin of U_D selected by instruments. IV, working through $S(Z, X)$, selects different slices of U_D and defines mean treatment effects for those slices.

In a model with a scalar and binary instrument with only two points in the support of $P(Z, X)$, the IV estimate and the overall LATE estimate are the same. When there are more than two distinct values of Z, an overall LATE (the standard IV estimator) can be estimated by a weighted average of the pairwise LATE parameters based on ordered values of the scalar instrument Z [26, 48]. However, Heckman et al. [23] showed that when a vector of instruments enter the choice model, the traditional IV method may produce misleading inferences since the IV estimate can be negative even if all the pairwise LATE estimates are positive. This is because the weights used to compute the overall LATE can be negative if the choice model is determined by a vector of instruments and the analyst uses only some of those instruments in the calculations [5, 23].

LATE is an interpretable parameter when the observed variation in the instrument defines the question for which the analyst seeks an answer, e.g., if the analyst has access to an instrument, Z, that takes two values (z and z') and the question he seeks to answer is precisely what happens when the instrument is changed from z_1 to z'. However, when the policy being analyzed does not conform closely to the instrument used, it is not always clear who the marginal patients associated with the policy are, and consequently, whether or not the marginal patients defined by LATE are those on which the clinical decision making should rely.

2.3.3 Marginal Treatment Effects—In order to address some of these limitations and to better understand the distribution of treatment effects in the population, we can use the Marginal Treatment Effect (MTE) first introduced by Björklund and Moffitt [7] (see also [16, 17, 19–22]). The MTE is the average gain to patients who are indifferent between receiving *treatment* 1 versus *treatment* 0 given X and Z. These are the patients at the margin as defined by X and Z. Formally, MTE can be defined by

$$\begin{aligned}\mathrm{MTE}(x,z)&=E(\Delta|X=x,Z=z,\Lambda=0)=E(\Delta|X=x,U_\Lambda=-\mu_\Lambda(z,x))\\&=\mu_1(x)-\mu_0(x)+E(U_1-U_0|U_\Lambda=-\mu_\Lambda(z,x))\\&=\mu_1(x)-\mu_0(x)+E(U_1-U_0|U_D=S(z,x)),\end{aligned}\tag{15}$$

where the last equality follows from the fact that $S(Z, X)$ is a monotonic transformation of the mean utility $\mu_V(Z, X)$ while U_D is a monotonic function of U_Λ. The mean conditional treatment effect at each level of U_D is the value of the MTE at that level of U_D. Evaluation of the MTE parameter at low values of U_D averages the outcome gain for those individuals whose unobservable characteristics make them less likely to undergo treatment, while evaluation of MTE parameter at high values of U_D gives the gain for those patients with unobservable characteristics which make them more likely to undergo treatment. For example, LATE is a weighted sum of all MTE within the margin at which LATE is identified. In the limit, as $\mu_V(z', x) \to \mu_V(z, x)$, LATE converges to MTE under standard regularity conditions.

An additional feature of MTE is that all mean treatment effects parameters, including the ATE, TT, and the IV effect, can be calculated from weighted averages of MTE. These weights can be obtained from the data [5, 22, 23]. For example, the ATE is the sum of all MTE across all distinct values of U_D, weighted equally (conditional on X). A more formal description of these weights is given below.

Equation (15) shows that the MTE is identified on the support of $S(Z, X)$, i.e., specific values of $S(Z, X)$ define the specific margin of indifference $U_D = u_D$. An average treatment effect at each level of U_D can be obtained by integrating MTE (x, u_D) over the distribution of X conditional on $U_D = u_D$. That is,

$$\begin{aligned}\mathrm{ATE}(u_D)&=E_{X|U_D=u_D}(\mathrm{MTE}(x,u_D))\\&=E_{X|U_D=u_D}\{(\mu_1(X)-\mu_0(X))+E(U_1-U_0|U_D=u_D)\}\end{aligned}\tag{16}$$

Additionally, by integrating these conditional ATEs over the distribution of U_D (which by construction is Uniform(0, 1)) we can obtain the (unconditional) Average Treatment Effect:

$$\mathrm{ATE}=E_{U_D}(\mathrm{ATE}(u_D))=E_{U_D}E_{X|U_D=u_D}(\mu_1(x)-\mu_0(x)+E(U_1-U_0|U_D=u_D))\tag{17}$$

Here, the last term in (17) drops out because $E_{U_D}E(U_1 - U_0|U_D = u_D) = E(U_1 - U_0) = 0$. Equation (17) suggests that the *weights* for the MTE(x, u_D) that yield the ATE can be constructed from the empirical joint distribution of $(X, S(Z, X))$ directly. Alternatively, since U_D is distributed as Uniform(0, 1), simply integrating ATE(u_D) over the full support of U_D yields ATE.

Obtaining the weights to estimate TT and the IV estimator is a bit more complicated than determining the weights for ATE, but they can be computed readily using the data at hand. Intuitively, for TT, the weights for MTE evaluated at high values of U_D are relatively larger than those evaluated at low values of U_D. This is because, by definition, larger values of U_D represent greater propensity to select treatment based on unobserved characteristics. The TT weights can be written as

$$\varpi_{TT}(x,u_D)=\frac{\Pr(S(Z)\le u_D|X=x,U_D=u_D)}{\int\int\Pr(D=1|X=x,U_D=u_D)du_D dF(X)} \tag{18}$$

Weights for other parameters such as the IV effect can be found in [23]. Using these weights one can reconstruct the traditional IV estimator based on the estimated distribution of MTEs. Such a reconstructed estimator will be denoted as the LIV-based IV estimator. It is used to show that a certain combination of MTEs can be used to explain the traditional IV results.

2.3.4 Estimators for MTEs and Other Mean Treatment-Effect Parameters—The method of local instrumental variable can be used to identify and estimate the MTE over the support of the propensity score, estimated using IVs in the choice equation, for selecting treatment [19, 22, 23]. In the LIV approach the outcome is modeled as a nonlinear function of all the X's and the whole propensity score, $P(Z)$, and interactions between them. What is important in this approach is to have a way to fully capture the nonlinearity of the outcome with respect to $P(Z)$. Now, if one takes the rate of change of the mean outcome with respect to $P(Z)$ evaluated at a particular value of $S(z, x) = 1 - P(z, x)$, one gets

$$\begin{aligned}\frac{\partial}{\partial P(z,x)}E(Y|Z=z,X=x)|_{1-P(z,x)=u_D}&=E((Y_1-Y_0)|X=x,u_D=1-P(z,x))\\&=\Delta+\frac{\partial K(P(z,x))}{\partial P(z,x)}=\mathrm{MTE}(x,u_D)\end{aligned} \tag{19}$$

where $K(P(z, x))$ is a differentiable function of $P(z, x)$. A formal derivation is given in the Appendix. Equation (19) shows that the key element for the estimation of MTE is the function $K(P(z, x))$. This function can be estimated using different econometric techniques, such as using flexible approximation to $K(P(z, x))$ based on a polynomial of the propensity score in a regression estimator or using fully non-parametric matching techniques. Specifically, in a regression context, (19) is implemented by regressing the outcome Y on all covariates, the estimated propensity score $\hat{P}(z, x)$, the interaction of the propensity score with all covariates, and a polynomial on the propensity score and then computing the partial derivative of the regression estimand with respect to the propensity score. Once MTE is estimated via LIV, the other mean treatment-effect parameters can also be estimated using different weighted averages of the estimated MTE, and these weights can be constructed from the data at hand. One of the limitations of LIV, however, is that it requires a sufficiently large sample size so as to identify the entire support of the propensity score. When this is not achieved, the LIV method can only produce upper or lower bounds to the mean treatment parameters, nevertheless, making this limitation explicit.

2.4 Comparative Effectiveness of Breast Cancer Treatments

2.4.1 Data—Our data come from the OPTIONS (Outcomes and Preferences in Older Women, Nationwide Survey) project [13]. The OPTIONS sample was designed to be representative of all female elderly Medicare beneficiaries (aged 67 or older) with newly diagnosed, early-stage breast cancer in Medicare's-fee-for service program between 1992 and 1994. Details of the specific exclusion criteria used can be found elsewhere [5, 14, 33]. These data provide a unique opportunity to analyze a large national sample of Medicare beneficiaries with confirmed local stage of breast cancer. The final sample consists of 2,517 patients of whom 1,813 patients had a MST and the remaining had BCSRT. The distribution of patient characteristics by treatment type is published elsewhere [33]. The outcome variable is 3-year survival rates, which was measured without any censoring.

The covariates that we control for are variables that are both measurable and theoretically predictive of survival. In addition to the treatment indicator, we include age at the time of surgery, cancer stage, Charlson co-morbidity index, race and urbanacity. These are also the typical covariates that are adjusted for in an RCT setting. The primary goal of the analysis is to estimate the distribution of marginal treatment effects (MTEs) and also to recover estimates for the average treatment effect (ATE) and the effect on the treated (TT) parameter on 3-year survival associated with BCSRT as compared to MST. Additionally, we compare these parameter estimates to the estimates produced by traditional IV analysis. We present two sets of analyses: (1) where confounder "age" is intentionally omitted and therefore this confounder contributes toward unobserved confounding and (2) where "Age" is included in the observed set of covariates.

The variables used as valid instruments include a regional dummy variable (NORTH) to represent regional variations in practice patterns, and a continuous variable that represents the Medicare physician fee differential (FEEDIF) between mastectomy and breast-conserving surgery calculated at the 3-digit zip-code level of the treating physician. NORTH represented a geographical variation in treatment selection, perhaps through a historical practice style, which is plausibly independent of underlying health, preferences and outcomes of the patients. In particular, women residing in the Northeast, Midwest and Pacific census divisions (represented by indicator NORTH) were more likely to receive BCSRT compared to MST. Medicare fees are assumed to be exogenous and independent of unobservable health of patients and preferences of patients and physicians because they were determined by a combination of the resource-based fee specified by the Medicare Fee Schedule, which is independent of any particular physician's or patient's characteristics, and the average historical Medicare payment in the geographic area. Further details and justification for these instruments are available in [13].

2.4.2 Methods—First, we estimate the propensity score of treatment choice as a function of all covariates and also the instruments NORTH and FEEDIF using a probit regression model. We use another probit regression model for the binary 3-year survival outcome (S). We implement the traditional IV estimator using a residual inclusion approach [45], where the first-stage (choice model) residuals are included as additional regressor in second-stage (outcome regression) estimation. For the LIV approach, we run the probit outcome regression on all covariates (X), the estimated propensity score ($\hat{p}$), the interaction of propensity score with all covariates, and a polynomial on the propensity score, $K(\hat{p}; d)$:

$$E(S)=\Phi^{-1}(\beta_0+X\cdot\beta_1+X\cdot\widehat{p}\cdot\beta_2+K(\widehat{p};d)) \tag{20}$$

The degree of polynomial, d, is selected based on both a likelihood-ratio test and a Wald-test of the joint test of significance for the polynomial coefficients. We use the derivative of the polynomial formulation as our LIV estimand, which is used to predict MTE(x, u_D). The predicted values of the propensity score allow us to define the values of U_D over which MTE can be identified [19]. The larger the support of the propensity score, the bigger the set over which MTE can be recovered.[5]

We reduce the dimensionality of X by using deciles of the estimated linear predictor in the LIV estimand that is only a function of the X and not the propensity scores.[6] We denote

[5]With parametric approaches, assumptions about functional form can estimate MTE over ranges of u_D that are not identified with our choice model and sample. This is not the case when non-parametric techniques are used instead.

[6]This is implemented by predicting $X\cdot\hat{\beta}_2$, where $\hat{\beta}_2$ corresponds to the estimated coefficients on the interaction term of X and $P(Z, X)$ in the LIV outcomes regression.

these deciles as η_q hereon, where $q = 1, 2, \ldots, 10$. Thus, using our coefficient estimates from the above regression (20) we estimate MTE(η_q, u_D) by varying u_D between 0 and 1 and using average predicted MTE(x, u_D) for each η_q. Note that MTE estimates using a value of $P(Z, X) = p$ are associated with $u_D = (1 - p)$. Using the empirical joint density of (η_q, u_D), which also represents the weights for MTE(η_q, u_D) required to calculate the *empirical* ATE (estimated over the observed common support), we estimate the MTE(u_D).

Next, we calculate the weights associated with ATE, TT and IV effect and use them to construct the respective treatment-effect estimates. Standard errors for MTE(u_D) and all the mean treatment-effect parameters are estimated via 1000 bootstrap replicates.

3 Results

3.1 With Age Omitted

Both instrumental variables are significant predictors of treatment choice ($p < 0.001$ for each). The left vertical panel of figures in Fig. 1 correspond to the analyses with confounder "age" omitted. Figure 1(a) illustrates the distribution of the predicted propensity score for choosing BCSRT separately for patients who chose BCSRT and those who chose MST. It also depicts the identified support where we find positive density of the propensity score for both treatment sub-samples. We cannot identify MTE over the entire (0, 1) support. Although we do find people near 0, there is essentially no mass close to 1. This means (unconditional) ATE is not identified in the sample without further assumptions [19, 22]. We, therefore, did not attempt to estimate am unconditional ATE but rather estimate an empirical ATE that was based on the margins of choices that we observe in the data [5].

The standard regression-based (IV-naive) estimate of the treatment effect was found to be 0.05 (Std. err. = 0.012) that was significant at 5% level. However, the standard IV-based estimate was −0.07 (Std. err. = 0.13). Although point estimate point toward harm caused by BCSRT over MST, it was not significant. The empirical estimation of the LIV estimand found that a cubic specification of the estimate propensity score was most appropriate for $K(\hat{p}; d)$.[7] The ATE(x) is displayed in Fig. 1(b) and shows no significant variation over η_q. The ATE(u_D) is displayed in Fig. 1(c) and shows considerable variation in treatment effects. In this figure, for u_D between 0.65 and 0.75 (higher values for U_D represents patients with latent characteristics that make then most likely to choose BCSRT compared to MST), the ATE(u) is significantly negative indicating that BCSRT is harmful for these margins. This effect disappears for the lower values of U_D, where ATE(u) estimates are close to zero and not significant. At higher values of U_D, ATE(u) estimates are positive favoring BCSRT over MST but these do not reach statistical significance. Since we could not identify the higher end of the support for $P(Z, X)$, we could not estimate MTE for the patients least likely to choose BCSRT (i.e. low u_D) based on their unobserved characteristics.

The estimated mean treatment-effect parameters are shown in Table 1. The LIV-based IV estimator produces as estimate of 0.10 (Std. err. = 0.10).[8] The unconditional TT and the empirical ATE estimates show a larger negative effect of BCSRT over MTE than the IV estimator, but were not significant.

[7]The Wald-F test for all higher order polynomials in a cubic specification was significant ($p = 0.02$). Compared to a quadratic specification, the likelihood-ratio test for the cubic specification was significant ($p = 0.007$). However, a quartic specification for our LIV estimand, compared to a cubic specification was marginally significant using the Wald-F test ($p = 0.05$) but not using the likelihood-ratio test ($p = 0.53$).

[8]This estimate is similar to the traditional IV estimator but not identical. We would not expect identical results between the two because the LIV based estimator considers the full interaction of treatment with observed confounders while the traditional IV estimator does not.

3.2 With Age not Omitted

Both instrumental variables are again significant predictors of treatment choice ($p < 0.001$ for each). Age categories also found to be significant predictors of choice. Specifically, 32% of 65–74 year olds choose BCSRT; compared to them, 75–79, 80–84 and 85+ year old patients choose BCSRT less by 5%pts ($p = 0.006$), 11%pts ($p < 0.001$) and 22%pts ($p < 0.001$), respectively.

The right vertical panel of figures in Fig. 1 correspond to the analyses with confounder "age" observed and not omitted from the regressions of the outcomes. The distribution of the predicted propensity score for choosing BCSRT separately for patients who chose BCSRT and those who chose MST (Fig. 1(a)) show similar margins of choice as those estimated without age.

The standard regression-based (IV-naive) estimate of the treatment effect was found to be 0.03 (Std. err. = 0.010) that was significant at 5% level. However, the standard IV-based estimate was −0.14 (Std. err. = 0.11). Although the point estimate was larger than when age was omitted and points toward harm caused by BCSRT over MST, it was not significant. The empirical estimation of the LIV estimand found that a linear specification of the estimate propensity score was most appropriate for $K(\hat{p}; d)$.[9] The ATE(x) is displayed in Fig. 1(b) and now shows significant variation over η_q. In fact the ATE(x) is significantly negative at the lower two deciles, η_1 and η_2. The ATE(u_D) is displayed in Fig. 1(c) and now shows considerably less variation in treatment effects as expected from a LIV estimand with a linear specification for $K(\hat{p}; d)$ within a probit regression model. The estimated ATE(u) is consistently flat over all values of U_D and does not reach statistical significance at any point.

The estimated mean treatment-effect parameters are shown in Table 1. As expected, since treatment effects were not found to vary over u_D. The LIV-based IV estimator and the empirical ATE estimator produce similar estimates of −0.17 (Std. err. = 0.13) (Table 1). The unconditional TT estimate is −0.15 (Std. err. = 0.13).

3.3 Discussion of Results

The differences in the estimated treatment-effect distributions and mean treatment-effect parameters between the two sets of analyses, where age is observed or not, highlight the role of treatment-effect heterogeneity over omitted variables in casual estimation using IV.

In order to better understand the differences, we computed the ATE for each age category in the same way we computed ATE(x) for any η_q. We found that the average treatment effect for age categories 64–74, 75–79, 80–84 and 85+ year olds are −0.07 (Std. err. = 0.12), −0.47 (0.22), −0.13 (0.24), and 0.018 (0.19), respectively. These estimates are in line with clinical intuition. MST does represent the most aggressive approach to remove the breast tumor. With BCSRT, one leaves open the possibility that the entire tumor was not removed from the body. Severity of diagnosed cancer increases with age. At younger ages, BCSRT may not be harmful compared to MST as the cancer is usually diagnosed at a very early stage. At older ages, the MST may not be beneficial as patients may die of many competing risks. However, somewhere in between, and in our analysis, from ages 74–79 years, the cancer is severe enough and competing risks of dying are low enough that MST provides significant survival benefits over BCSRT ($p = 0.03$).

[9]Neither of the higher order specification (quadratic, cubic or quartic) specification passed both the Wald-F and the likelihood-ratio tests.

The IV estimator estimates an effect for the margin in the population that are induced to choose BCSRT due to levels of the instruments. This IV effect is usually conditional on a specific level of unobserved confounder. When age is omitted, the heterogeneity in treatment effects over age manifests as unobserved treatment-effect heterogeneity. Consequently, the IV estimator identifies a local effect corresponding to the specific level of unobserved confounder that the IVs hold constant. That is why the IV effect came out to be different than ATE or TT in our application but the difference was not statistically significant. The LIV method, however, divulged the distribution of ATE(u) in the population and identified certain margins where BSCRT was harmful compared to MST. Such information can spur further research to study risk factors that may be driving such negative BSCRT in these margins of the population.

In our application, such a risk factor was age as evident from our second set of analysis where age was included as an observed confounder. Treatment effect no longer varied over unobserved confounder but showed significant variability over observed confounder, especially age. In this case, the IV estimator produces a consistent estimate of ATE. Although not necessary, in our application the effect on the treated also seems fairly similar to the average treatment effect in this population. This is because the marginal distribution of age categories 64–74, 75–79, 80–84 and 85+ year olds in the population is 0.59, 0.24, 0.12, and 0.05, while the marginal distribution of the same conditional on BCSRT choice is 0.66, 0.22, 0.09, and 0.02. For patients 75–79 years old, where the treatment effect is largest and significant, the proportion among BCSRT choosers is the same as that in the general patient population.

Clinical trial results and average IV results fails to divulge such heterogeneity. It is intuitive to assume that patients who enrol in clinical trials may not have strong preference for either BCSRT or MST. If they do, they would directly receive those treatments instead of enrolling in a clinical trial. In our analysis, if we look at the margins of choice given by u_D, which represents propensity to select treatment based on unobserved confounders, and focus on u_D close to 0.5, we can see that the average treatment effect at those margins are close to zero. It is quite possible that the clinical trials are estimating the effect only at these margins. However, a confirmatory analysis for this hypothesis is beyond the scope of this paper and is left for future work.

4 Conclusions

These results have many implications for future CER studies. During 1990–1992, from when this dataset belong, both the average effect and the effect on the treated for BCSRT versus MST appear to be negative. Although these estimates do not reach statistical significance, the LIV approach reveal distinct margins where BCSRT produces significant negative effects on survival compared to MST. However, clinical trials results and other traditional IV analyses that concluded that on average BCSRT has equivalent effect on survival as MST may have been influential in the diffusion of BSCRT over the years. It is plausible that a similar analysis with more recent data may reveal that effect on the treated for BCSRT as compared to MST may have exacerbated. Therefore, estimating and correctly interpreting treatment-effect heterogeneity appears to be critical for any CER study.

Our analyses also have implication about how cautious we have to be in generating comparative effectiveness information. Generating internally valid estimates of treatment effects (such as those in randomized clinical trials) is not sufficient for realizing the anticipated goals of CER. Understanding the generalizability of such estimates and promoting research in exploring the full distribution of treatment effects in the population will be crucial for the purpose of effective translation of CER results to practice.

The local instrumental variable provides a novel and important approach to explore observed and unobserved heterogeneity in comparative treatment effects.

Acknowledgments

The author is grateful for helpful comments from an anonymous reviewer and is also grateful to Daniel Polsky at the University of Pennsylvania and the OPTIONS team for allowing to use their breast cancer dataset. The author would like to acknowledge financial support from a National Cancer Institute research grant 1 RC4 CA155809-01 (PI Basu).

References

1. Amemiya T. The non-linear two-stage least squares estimator. J Econom. 1974:105–110.
2. Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996; 91:444–455.
3. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005; 61:962–972. [PubMed: 16401269]
4. Basu A. Individualization at the heart of comparative effectiveness research: the time for i-CER has come. Med Decis Mak. 2009; 29(6):N9–N11.
5. Basu A, Heckman J, Navarro-Lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Econ. 2007; 16(11):1133–1157. [PubMed: 17910109]
6. Basu A, Philipson T. Impact of comparative effectiveness research on health and healthcare spending. NBER working paper No. w15633. 2010
7. Björklund A, Moffitt R. The estimation of wage gains and welfare gains in self-selection. Rev Econ Stat. 1987; 69(1):42–49.
8. Blundell, R.; Powell, J. Endogeneity in nonparametric and semiparametric regression models. In: Hansen, L.; Dewatripont, M.; Turnovsky, SJ., editors. Advances in economics and econometrics. Cambridge University Press; Cambridge: 2003. p. 312-357.
9. Brooks JM, Chrischilles E, Scott S, Chen-Hardee S. Was lumpectomy underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa. Health Serv Res. 2003; 38(6):1385–1402. Part I. [PubMed: 14727779]
10. Earle CE, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrument variable and propensity analysis. J Clin Oncol. 2001; 19(4):1064–1070. [PubMed: 11181670]
11. Fisher B, Bauer M, Margolese R, et al. Five-year results of a randomized clinical trial comparing total mastectomy and segmental mastectomy with or without radiation in the treatment of breast cancer. N Engl J Med. 1985; 312:665–673. [PubMed: 3883167]
12. Franklin BA. Lessons learned from the COURAGE trial: generalizability, limitations, and implications. Prev Cardiol. 2008; 11(1):5–7. [PubMed: 18174784]
13. Hadley J, Mitchell JM, Mandelblatt J. Medicare fees and small area variations in the treatment of localized breast cancer. N Engl J Med. 1992; 52:334–360.
14. Hadley J, Polsky D, Mandelblatt JS, Mitchell JM, Weeks JC, Wang Q, Hwang Y-T. OPTIONS Research Team. An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a Medicare population. Health Econ. 2003; 12:171–186. [PubMed: 12605463]
15. Heckman JJ. Comments on angrist, imbens, and rubin: identification of causal effects using instrumental variables. J Am Stat Assoc. 1996; 91:434.
16. Heckman J. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. J Hum Resour. 1997; 32(3):441–462.
17. Heckman J. Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture. J Polit Econ. 2001; 109(4):673–748.
18. Heckman J, Honore B. The empirical content of the Roy model. Econometrica. 1990; 58:1121–1149.

19. Heckman JJ, Vytlacil EJ. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proc Natl Acad Sci USA. 1999; 96(8):4730–4734. [PubMed: 10200330]
20. Heckman, J.; Vytlacil, E. Econometric evaluation of social programs. In: Heckman, J.; Leamer, E., editors. Handbook of econometrics. Vol. 6. Elsevier; Amsterdam: 2005.
21. Heckman J, Vytlacil E. Structural equations, treatment effects and econometric policy evaluation. Econometrica. 2005; 73(3):669–738.
22. Heckman, J.; Vytlacil, EJ. Econometric evaluation of social programs, part II: using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In: Heckman, J.; Leamer, E., editors. Handbook of econometrics. Vol. 6B. Elsevier; Amsterdam: 2007.
23. Heckman JJ, Urzua S, Vytlacil E. Understanding instrumental variables in models with essential heterogeneity. Rev Econ Stat. 2006; 88(3):389–432.
24. Hernan MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008; 19:766–779. [PubMed: 18854702]
25. Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. Ann Intern Med. 2002; 137:273–84. [PubMed: 12186518]
26. Imbens G, Angrist J. Identification and estimation of local average treatment effects. Econometrica. 1994; 62(2):467–475.
27. McClellan M, McNeil B, Newhouse J. Does more intensive treatment of acute myocardial infarction reduce mortality? JAMA. 1994; 272(11):859–866. [PubMed: 8078163]
28. McFadden, D. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P., editor. Frontiers in econometrics. Academic Press; New York: 1973.
29. McFadden, D. Econometric models of probabilistic choice. In: Manski, CF.; McFadden, D., editors. Structural analysis of discrete data with econometric applications. MIT Press; Cambridge: 1981.
30. Mullahy J. Instrumental variable estimation of count data models: applications to models of cigarette smoking behavior. Rev Econ Stat. 1997; 79:586–593.
31. Quandt RE. A new approach to estimating switching regression. J Am Stat Assoc. 1972; 67(338): 306–310.
32. Quandt RE. The estimation of parameters of a linear regression system obeying two separate regimes. J Am Stat Assoc. 1958; 53(284):873–880.
33. Polsky D, Mandelblatt JS, Weeks J, Venditti L, Hwang YT, Glick HA, Hadley J, Schulman KA. Economic evaluation of breast cancer treatment: considering the value of patient choice. J Clin Oncol. 2003; 21:1139–1146. [PubMed: 12637482]
34. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc. 1995; 90:106–121.
35. 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. 2002; 288:321–33. [PubMed: 12117397]
36. Rosenbaum PR, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.
37. Roy AD. Some thoughts on the distribution of earnings. Oxf Econ Pap. 1951; 3:135–146.
38. Rubin D. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics. 1973; 29:185–203.
39. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med. 1997; 127:757–763. [PubMed: 9382394]
40. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Stat Assoc. 1994; 94:1096–1120. (with rejoinder, 1135–1146).

41. Steering Committee on Clinical Practice Guidelines for the Care and Treatment of Breast Cancer. Mastectomy or lumpectomy? The choice of operation for clinical stages I and II breast cancer. Can Med Assoc J. 1998; 158(Suppl 3):S15–S21. [PubMed: 9484274]
42. Stock JH, Trebbi F. Who invented instrumental variable regression? J Econ Perspect. 2003; 17(3): 177–194.
43. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA. 2007; 297:278–285. [PubMed: 17227979]
44. Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration—a simulation study. Am J Epidemiol. 2007; 165:1110–1118. [PubMed: 17395595]
45. Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. J Health Econ. 2008; 27(3):531–543. [PubMed: 18192044]
46. Vanness, DJ.; Mullahy, J. Perspectives on mean-based evaluation of health care. In: Jones, A., editor. The Elgar companion to health economics. Edward Elgar Publishing; Cheltenham: 2006.
47. Virnig BA, et al. Increased use of breast-conserving surgery: preferred treatment or failure to provide adequate local therapy? Breast Cancer Res Treat. 2007; 106(Supp 1) Abstract 4065.
48. Yitzhaki, S. Working paper. Vol. 217. Department of Economics, Hebrew University; 1989. On using linear regression in welfare economics.

Appendix: Derivations for (19)

$$\begin{aligned}
E(Y|Z=z,X=x) &= E(DY_1+(1-D)Y_0|Z=z,X=x)\\
&= E(Y_0|X=x)+E(D(Y_1-Y_0)|Z=z,X=x)\\
&= E(Y_0|X=x)+\Pr(D=1|Z=z,X=x)\times E((Y_1-Y_0)|D=1,X=x)\\
&= E(Y_0|X=x)+\Pr(D=1|Z=z,X=x)\times \frac{\int_{S(z,x)=1-P(z,x)}^{1} E((Y_1-Y_0)|U_D=u,X=x)du}{\int_{S(z,x)=1-P(z,x)}^{1} du}\\
&= E(Y_0|X=x)+\int_{S(z,x)=1-P(z,x)}^{1} E((Y_1-Y_0)|U_D=u,X=x)du,
\end{aligned}$$

where the last equality follows as $D = (U_D > S(z, x))$ and therefore,

$$\Pr(D=1|Z=z,X=x)=\int_{S(z,x)=1-P(z,x)}^{1} du$$

Now, if we take the rate of change of the mean outcome with respect to the probability of receiving treatment evaluated at a particular value of $S(z, x) = 1 - P(z, x)$:

$$\begin{aligned}
\frac{\partial}{\partial P(z,x)}E(Y|Z=z,X=x)|_{1-P(z,x)=u_D} &= E((Y_1-Y_0)|X=x,u_D=1-P(z,x))\\
&= \mathrm{MTE}(x,u_D)
\end{aligned}$$

The formal proof of consistency for this estimator can be found in [23].

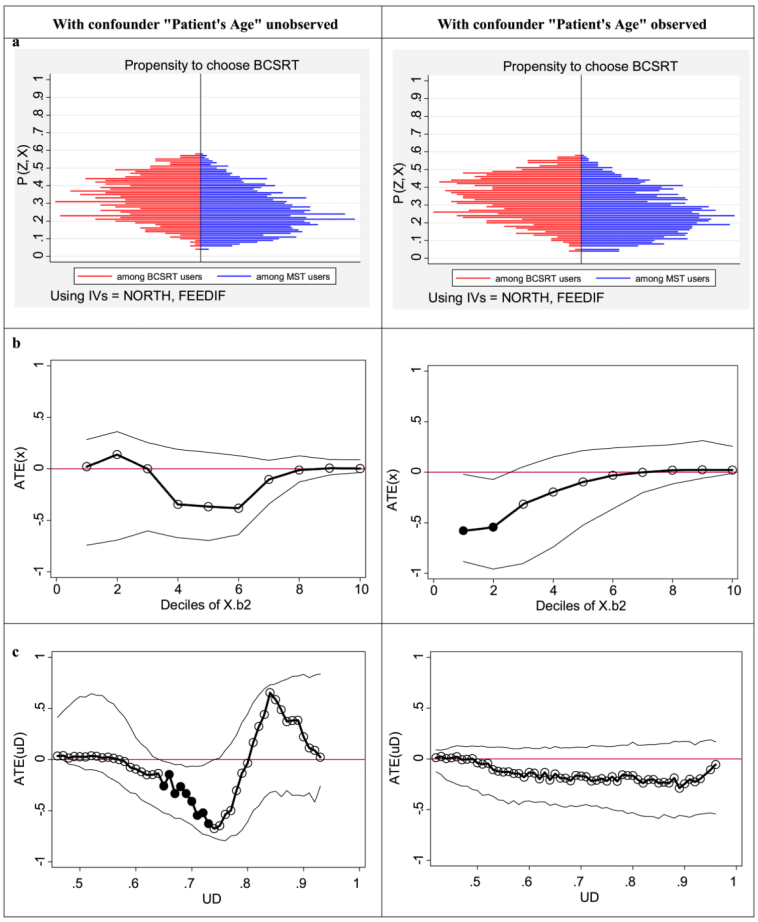

Fig. 1.

(**a**) Estimated propensity for choosing BCSRT among BCSRT and MST receivers. (**b**) Heterogeneity in treatment effects across deciles (η_q, $q = 1, \ldots, 10$) of $X\beta_2$. (**c**) Heterogeneity in treatment effects across U_D, propensity to select BCSRT based on unobserved confounders. (*Solid circles* in (**b**) and (**c**) represent treatment effects that are significant at 5% level)

Table 1

Mean treatment effects

Treatment effects	With age omitted	With age observed
	Mean (se)	Mean (se)
Naive estimate	0.05 (0.012)	0.03 (0.01)
IV estimate	−0.07 (0.13)	−0.14 (0.11)
LIV-based estimates		
IV effect[a]	−0.10 (0.10)	−0.17 (0.13)
Empirical ATE	−0.15 (0.09)	−0.17 (0.13)
TT	−0.12 (0.09)	−0.15 (0.13)

[a]This is the reconstructed IV estimator based on the estimated distribution of MTEs and the IV weights [23]. It is used to show that a certain combination of MTEs can be used to explain the traditional IV results