
Methods

Who Is the Marginal Patient? Understanding Instrumental Variables Estimates of Treatment Effects

Katherine M. Harris and Dahlia K. Remler

Objective. To clarify the issues of generalizability arising from the use of instrumental variable (IV) methods to estimate treatment effects in nonexperimental medical outcome studies.

Data Source. We generate Monte Carlo data designed to resemble typical data sets where detailed health status information is unavailable and the treatment assignment process is unobserved. The model used to generate our data makes the realistic assumption that unobservable health status characteristics of patients influence the treatment assignment process and the effectiveness of treatment.

Study Design. We use Monte Carlo data to illustrate the circumstances where IV estimates generalize to an unobservable patient subpopulation and those where IV estimates generalize to the entire patient population represented by the sample used in the analysis. We also simulate the effect of two policy changes that affect practice patterns. Further, we show that IV estimates are useful for predicting the effect of these changes on treatment effectiveness when the subpopulation to which the IV estimate refers is the same or very similar to the population whose treatment status is affected by the policy change.

Conclusions. Health services researchers cannot take for granted that IV estimates generalize to the same population represented by the sample used for analysis. Instead, researchers must rely on their knowledge of clinical practice and theory regarding the treatment assignment process in interpreting their results and in predicting the effect of changes in practice patterns.

Key Words. Instrumental variables, nonexperimental research, medical outcomes

The goal of outcomes research is to determine the effects of medical interventions on the health and well-being of patients. Ethical, financial, and practical considerations limit the widespread use of randomized experiments in medical outcome studies. Consequently, the use of nonexperimental data for medical outcome studies is growing dramatically with the increased availability of computerized medical data. Without random treatment assignment,

it is not possible to measure effectiveness by simply comparing the health outcomes of treated patients with those of untreated patients (Andersen 1994; Holland 1986; Institute of Medicine [IOM] 1985). This is because treatment decisions depend on unobserved characteristics of patients and providers that influence outcomes whether patients receive treatment or not. Because it is difficult, if not impossible, for researchers to measure and control for all of the factors that influence both treatments and outcomes, it is difficult to know whether observed outcomes are due to treatment or due to unobserved patient differences.

Interest is growing among health services researchers in the use of instrumental variable (IV) methods¹ as a substitute for random treatment assignment in nonexperimental medical outcome studies. An IV is an observable variable that can be used in lieu of a coin-flip in assigning patients to the treatment of research interest.² Appropriately applied, IV methods represent a valuable tool for inferring the effect of medical treatments on a range of potential outcomes in situations where randomization is unethical or impractical. Researchers using IV methods need to assess the validity of two types of assumptions: (1) those required for an IV to act as a randomizing variable and (2) those governing the generalizability of the estimated parameters. Until recently, the plausibility of assumptions required for IV validity have received more attention than those required for the ability to generalize.

IV methods have been widely used in the literature on evaluation to measure the effects of social programs (e.g., see Angrist 1995 and references therein; Currie and Gruber 1996; Levitt 1996.) Typically, this literature has assumed that the effect of program participation or "treatment" is the same for everyone. If treatment effects are homogenous, then the generalizability of IV estimates is straightforward. Only recently has attention in the statistical and econometric literatures focused on the issue of generalizability in situations where the effect of treatment is assumed to vary across individuals (Heckman 1997; Angrist, Imbens, and Rubin 1996; Moffitt 1996; Meyer 1995; Imbens and Angrist 1994).

These studies show that when treatment effects are heterogeneous, IV estimates are determined by variation in the outcome of interest among a

Katherine M. Harris, Ph.D. is an Assistant Professor in the Institute for Health Services Research, Tulane University Medical Center. Address correspondence and requests for reprints to Dahlia Remler, Ph.D., Assistant Professor, Institute for Health Services Research, Tulane Medical Center, 1430 Tulane Avenue, New Orleans, LA 70112-2699. This article, submitted to *Health Services Research* on July 18, 1997, was revised and accepted for publication on January 23, 1998.

sample subgroup whose treatment status depends on the value of the IV. In this literature, these sample members are referred to as the “marginal” subgroup. This subgroup is analogous to persons in experimental studies who meet inclusion criteria and who comply with random treatment assignment but who are not necessarily typical of the population for whom the treatment is designed (Angrist, Imbens, and Rubin 1996; Robins and Greenland 1996). In nonexperimental settings, however, the marginal subgroup is unlikely to be observable with readily available data. Thus, when treatment effects are heterogeneous, researchers cannot take for granted that IV estimates of treatment effects are determined by all sample members used in an analysis. As a result, it may not be possible to draw inferences about the effect of treatment in the population represented by a particular sample of people.

This is relevant to health services researchers, because in clinical settings heterogeneous treatment effects are common. For example, the effect of thrombolytic drugs on the health outcomes of stroke victims depends critically on the underlying health status of these patients. Those who suffer from blood clots often return to their original functional status, while persons with brain hemorrhage have their conditions worsened (von Kummer, Allen, Holle, et al. 1997; Quality Standards Subcommittee of the American Academy of Neurology 1996). Alternatively, the effect of a chemotherapy series on mortality depends on the stage at which the particular cancer is diagnosed and treated. By contrast, clinical examples of homogenous treatment effects are less obvious. One example might be the effect of appendectomies in returning virtually all patients with appendicitis to their prior state.

In circumstances where treatment effects are heterogeneous, health services researchers who use IV methods must rely on their knowledge of clinical practice and on theory about the treatment assignment process in generalizing from their results. Moreover, researchers using IV estimates to predict the effect of policy interventions must be able to justify on theoretical grounds that the IV estimates generalize to the same patient population whose treatment status will be affected by the policy change.

Confusion over the generalizability of IV estimates is evident in the responses to a widely read study by McClellan, Newhouse, and McNeil (1994) that uses a variable related to travel distance³ as an IV intended to measure the effect of invasive cardiac procedures on a subgroup of Medicare heart attack patients.⁴ The authors are careful to note that their finding of a lack of improvement applies only to a subgroup of marginal patients for whom receiving treatment depends on their distance to a hospital equipped for invasive cardiac procedures. Nonetheless, an accompanying editorial

by Gould (1994) and subsequent letters to the editor (Davis 1995; Gross and Shua-Haim 1995; Kuller and Detre 1995) illustrate confusion about generalizability. The implicit assumption is that the findings generalize to the entire population of Medicare heart attack patients, rather than to a "marginal" subpopulation.

The purpose of this article is to clarify issues surrounding the generalizability of treatment effects measured with IVs. We use Monte Carlo methods to generate hypothetical data resembling that typically used in nonexperimental medical outcome studies where detailed health status information and the treatment assignment process are unobserved. Monte Carlo methods offer two advantages. First, they allow us to focus on generalizability issues without having to worry about the validity of our IVs because they are valid by construction. Second, we can use knowledge of the (otherwise unobservable) data-generating process to compare treatment effects estimated with IVs to their true population values. Our illustrations show how two different valid IVs can produce different treatment effect estimates because they refer to different subpopulations, and how failure to consider these different subpopulations can form the basis of misleading inferences.

AVERAGE TREATMENT EFFECTS

This section introduces the concept of an "average treatment effect" as a summary measure of treatment benefits in situations where patients vary in their responses to medical treatment. We start with a simple two-equation model of the effect of medical treatment on health outcomes and the treatment assignment process. The assumption that the effect of treatment varies across individuals on the basis of their health status distinguishes this model from that underlying many experimental and nonexperimental studies in both the biomedical and program evaluation literatures. This assumption makes the model considerably more realistic from the point of view of medical outcomes research, and it has important implications for the generalizability of treatment effect estimates.

For simplicity, our discussion does not include the effect of observable characteristics on health outcomes. One can think of the model as being applied separately to groups of patients in which each person has identical observable characteristics. In this model, patient i 's health outcome can be written

$$y_i = \beta(h_i)d_i + g(h_i) + \varepsilon_i \quad I = 1, N \quad (1)$$

where h_i denotes unobservable health status; d_i is a dichotomous variable that takes the value one if the patient is treated and zero otherwise; $\beta(h_i) + g(h_i)$ is the expected health outcome if the patient receives the treatment; $g(h_i)$ is the expected health outcome if the patient does not receive the treatment; and ε_i represents the effect of other unobserved factors unrelated to health status. The effect of treatment for each individual is the difference between health outcomes in the treated and untreated state, $\beta(h_i)$. If treatment effects are homogenous, then $\beta(h_i) = \beta$ for everyone. If treatment effects are heterogeneous, then $\beta(h_i)$ is different for everyone.

Next, the probability that patient i receives treatment can be written

$$P(d_i = 1) = f(h_i) + z_i \quad (2)$$

where $f(h_i)$ represents health status characteristics that determine treatment assignment, and z_i represents factors uncorrelated with health status that have a nontrivial impact on the probability of receiving treatment. In this model, we assume that researchers cannot observe all of the elements of h_i and ε_i .

If it were possible to observe each patient in the treated and untreated state, we could simply measure the values of the treatment effects, $\beta(h_i)$, for every sample person and report each value individually or report a variety of sample statistics describing the distribution of these values. Because this is not possible, it is necessary to devise other strategies for obtaining information about treatment effectiveness.

Randomization can solve the problem of not being able to observe patients in the unrealized state (Heckman and Smith 1995; Holland 1986; IOM 1985; Rubin 1978, 1974). Random treatment assignment helps to assure that on average the unobservable characteristics of patients who receive treatment are the same as those of patients who do not. In experimental settings, researchers strive to eliminate the effect of health status on the treatment assignment process shown in Equation 2 by randomly generating (perhaps in the form of a coin-flip) values of z_i such that they are uncorrelated with health status and then assigning subjects to treatment and control groups on the basis of its value.

In this way, randomization makes it safe to infer that the difference in mean outcomes between treated and untreated patients, $\bar{y}_{\text{treated}} - \bar{y}_{\text{untreated}}$ is due to the treatment of research interest and not to other unobservable factors. When treatment effects are heterogeneous, this difference in mean outcomes is equal to β . In the heterogeneous case considered in this article, this difference represents an average over all of the individual treatment effects. This average is called an "average treatment effect" and is referred to as $\bar{\beta}$.

In the absence of randomization, however, we cannot expect treated and untreated patients to be alike on average. If unobservable characteristics influence outcomes and treatment assignment as in Equations 1 and 2, then the simple difference between $\bar{y}_{\text{treated}}$ and $\bar{y}_{\text{untreated}}$ does not yield a number that can be interpreted as an average treatment effect. This problem is often referred to as "selection bias." In situations where physicians treat the healthiest patients, who would have had better outcomes even without treatment, the estimated average treatment effect overstates the benefits of treatment. By contrast, in situations where physicians reserve treatment for the frailest patients, who would have had worse outcomes without treatment, the estimated average treatment effect understates the benefits of treatment.

In some nonexperimental settings, it may be possible to identify one or more naturally occurring z_i s that influence treatment status and are otherwise uncorrelated with health status. When this is the case, it is possible to estimate a parameter that represents the average effect of treatment among the subgroup of patients in the sample for whom the IV determines treatment assignment.

VALIDITY OF INSTRUMENTAL VARIABLES

The validity of the estimated treatment effect depends critically on whether the IV really acts to determine treatment status in a random fashion and has no independent effect on health outcomes. Four conditions are required for a valid IV.^{5,6,7} First, the IV must influence the probability of receiving treatment. In other words, we must be able to think of the IV as a z_i in Equation 2. Use of distance by McClellan, Newhouse, and McNeil (1994) as an IV is intuitively appealing because people having heart attacks are generally taken to the nearest hospital. The authors assume that heart attack patients who live closer to hospitals equipped to perform invasive cardiac procedures are more likely to receive invasive treatment than those who live far from equipped hospitals. It is possible to assess the validity of this assumption with observable data by testing whether or not an IV is a statistically significant predictor of the receipt of treatment.

Second, the IV must have no independent effect on outcomes. Thus, McClellan, Newhouse, and McNeil assume that distance influences health outcomes only through its effect on treatment status. In other words, it must be the case that z_i does not belong in Equation 1 and is not a proxy for any variable that should be included in Equation 1 but is omitted. This is called the exclusion restriction and cannot be tested readily with observational data.⁸

Third, the IV must be exogenous or “ignorable.” Health outcomes, or any other variable that belongs in Equation 1, must not influence the value of the IV. In other words, z_i in Equation 2 must not depend on health status, h_i ; other characteristics, ε_i ; or outcome, y_i . For example, distance would not be a valid IV if people with poor health status chose to live closer to hospitals equipped to perform invasive procedures.

These last two assumptions assure that there are no differences between groups formed on the basis of the value of the IV that could be confounded with the effect of treatment. In order to assess their validity, McClellan, Newhouse, and McNeil split their sample into two groups, those who are near to an equipped hospital and those who are far from one, and compare the two groups on the basis of observable characteristics. The authors show that the near and far groups are reasonably similar on all observable characteristics. However, the authors are not able to rule out the possibility that the two groups differ on the basis of characteristics that are not observable in their data.

Finally, the relationship between the IV and the probability of receiving the treatment must be “monotonic” when treatment effects are heterogeneous and depend on the same unobserved characteristics that influence treatment assignment.⁹ In the case of distance, any particular patient’s chances of receiving the invasive treatment must increase as the patient moves closer to an equipped hospital, holding other characteristics (e.g., severity or hospital staffing) constant. Similarly, any particular patient’s chances of receiving the invasive treatment must decrease as the patient moves farther from equipped hospitals. In other words, the monotonicity assumption rules out the possibility that patients’ chances of receiving treatment decrease as they move closer to equipped hospitals. This assumption is also not testable with observable data.

In addition to distance, IVs that have been used in medically related studies include state laws determining Medicaid eligibility and malpractice (e.g., Currie and Gruber 1996; Kessler and McClellan 1996). The area variations literature suggests that medically identical patients receive different treatments in different regions (Wennberg 1987; Wennberg and Gittelsohn 1982). Guadagnoli, Hauptmann, Ayanian, et al. (1995) show differences in treatment of heart attacks between New York and Texas. This suggests that geographic region can serve as an IV. On the other hand, it is also reasonable to assert that region belongs in the outcome Equation 1 as well as in the treatment assignment Equation 2 if region is a proxy for unobserved health status and therefore that region is not a valid instrument.

IV ESTIMATES OF TREATMENT EFFECTS

In applications where it is possible to identify a valid IV, patient characteristics continue to play a role in determining treatment status. Knowledge of clinical practice suggests that for some subgroups of patients even a valid IV plays no role in determining treatment assignment. In the context of the McClellan, Newhouse, and McNeil study, for example, physicians may view some patients as too frail to receive treatment even if they live near an equipped hospital. Alternatively, physicians may view other patients as such good candidates that they receive treatment even if they live far from an equipped hospital. Thus, the IV determines treatment status only for those patients whose health status does not place them in one of these two groups. Imbens and Angrist (1994) show that the average treatment effect measured by the IV refers only to the marginal subgroup of patients whose treatment status was determined by the value of the IV.

Experimental medical outcome studies also measure an average treatment effect that generalizes to a subpopulation of patients. For ethical reasons, this may be a subgroup that has exhausted all other treatment options. For statistical reasons, this may be the subgroup among whom researchers think they are most likely to observe an effect. Just as researchers use care in generalizing experimental findings to broader patient populations, researchers using IV methods must use care in generalizing treatment effect estimates beyond the subpopulation of marginal patients. In nonexperimental studies, however, the lack of formal inclusion criteria and the inability to monitor compliance with treatment assignment lead to an inability to observe the marginal subgroup, further complicating the assessment of generalizability of IV estimates.¹⁰

Figure 1 helps to illustrate these issues. We borrow from use of distance as an IV by McClellan, Newhouse, and McNeil and let z_i represent distance to a hospital equipped to perform invasive procedures on heart attack patients. In our illustration, distance on the vertical axis takes on two values: $z_i^{dist} = 1$ if the patient lives close to an equipped hospital, and $z_i^{dist} = 0$ if the patient does not live close to an equipped hospital. Patient health status, h_i on the horizontal axis, takes on values between -1 and 1 where larger values represent better health. We make the line separating the near and far groups parallel to the health status axis to indicate that both groups have the same health status on average. In Figure 1, dotted regions represent patients who do not receive treatment, and diagonal lines represent patients who do receive treatment. In this illustration, we imagine that physicians never perform

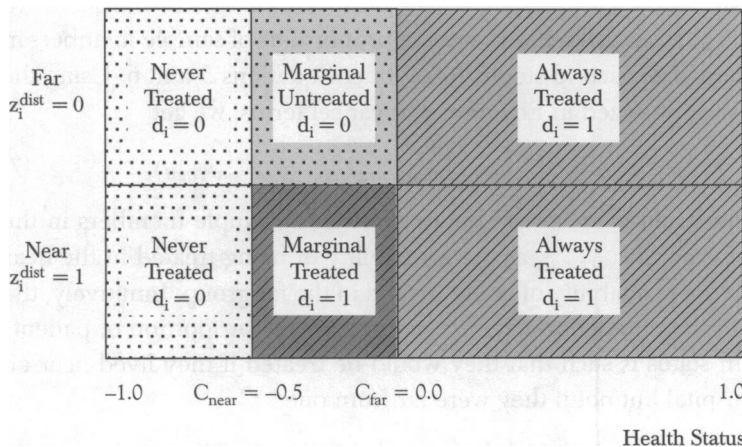
invasive procedures on patients in poor health status for fear of worsening outcomes and that they always perform invasive procedures on the most healthy patients in anticipation of improving outcomes.

Near patients receive treatment if their health status is greater than the cutoff, $C_{near} = -0.5$. Far patients receive treatment if their health status is greater than $C_{far} = -0.0$. Patients with health status greater than C_{far} are healthy enough to receive treatment always, regardless of distance. Patients with health status less than C_{near} are frail enough to receive treatment never, regardless of distance. Patients with health status between $C_{far} = -0.5$ and $C_{near} = 0.0$, represented by the shaded middle region of Figure 1, receive treatment if near but not if far. We imagine that physicians are uncertain about the benefits of treatment for this group. McClellan, Newhouse, and McNeil refer to members of this subgroup as “marginal” patients.

If distance is a valid IV, then we expect mean health outcomes to be the same on average within members of the always and never groups when the sample is divided into near and far groups. Specifically, in any sample we expect

$$\bar{y}_{near,never} - \bar{y}_{far,never} = 0 \text{ and } \bar{y}_{near,always} - \bar{y}_{far,always} = 0. \quad (3)$$

Figure 1: Treatment Status by Health Status and Value of the Instrumental Variable Representing Distance to a Treatment Facility



Note: Shaded regions represent marginal patients, regions with diagonal pattern represent treated patients, and regions with dot pattern represent untreated patients.

This is because both groups have the same treatment status and distance is unrelated to other factors that influence outcomes. By contrast, we expect mean health outcomes to be different within the marginal group when the sample is divided into near and far. This is because, in our example, the near marginal patients receive the treatment and the far marginal patients do not. Thus, any differences in outcomes resulting from the treatment can be detected only among marginal patients.

If it were possible to observe the marginal patients directly, their average treatment effect could be calculated by dividing the marginal patients into near and far and taking the difference in mean outcomes between those subgroups.¹¹

$$\hat{\beta}_{IV} = \bar{y}_{near,marginal} - \bar{y}_{far,marginal} \quad (4)$$

While we cannot observe the marginal group directly, it is possible to rewrite $\hat{\beta}_{IV}$ in terms of observable data (the mean outcomes for near and far groups and the proportion of patients who are in the marginal subgroup) using the assumptions required for a valid IV. The first step is to decompose the difference in mean outcomes between near and far groups into the contributions due to the marginal, never, and always subgroups,

$$\bar{y}_{near} = \pi_{marginal} \bar{y}_{near,marginal} + \pi_{never} \bar{y}_{near,never} + \pi_{always} \bar{y}_{near,always} \quad (5)$$

$$\bar{y}_{far} = \pi_{marginal} \bar{y}_{far,marginal} + \pi_{never} \bar{y}_{far,never} + \pi_{always} \bar{y}_{far,always} \quad (6)$$

where $\pi_{marginal}$, π_{never} , and π_{always} are the proportions of sample members in each subgroup. The second step is to subtract Equations 5 and 6. Using the assumptions summarized in Equation 3 to cancel terms, we get

$$\bar{y}_{near} - \bar{y}_{far} = \pi_{marginal}(\bar{y}_{near,marginal} - \bar{y}_{far,marginal}). \quad (7)$$

The third step is to rewrite the proportion of sample members in the marginal subgroup, $\pi_{marginal}$, as the probability of being treated in the near group minus the probability of being treated in the far group. Intuitively, the proportion of the population who are marginal is the proportion of patients whose health status is such that they would be treated if they lived near an equipped hospital but not if they were far from one.

$$\pi_{marginal} = P(d_i = 1 | z_i^{dist} = near) - P(d_i = 1 | z_i^{dist} = far) \quad (8)$$

The final step is to rearrange Equation 7 using the definition in Equation 8. The result is the IV estimate written in terms of observable data:

$$\hat{\beta}_{IV} = \frac{\bar{y}_{near} - \bar{y}_{far}}{P(d_i = 1|z_i^{dist} = near) - P(d_i = 1|z_i^{dist} = far)}. \quad (9)$$

If all patients were marginal, the denominator would be equal to one and $\hat{\beta}_{IV}$ would simply equal the mean difference between the near and far groups. If marginal patients comprise a small fraction of the near and far groups, then even a small difference in mean outcomes between near and far groups could create a substantial average treatment effect. This is due to the scaling effect of a small denominator.

In nonexperimental settings, researchers rarely have the opportunity to observe directly the process by which patients are assigned to treatment. Thus, in most cases, it is not possible to identify the marginal subgroup of patients whose average treatment effect is measured by the IV. In order to interpret the IV estimate, researchers must use their understanding of clinical practice to identify which patients are marginal.

Two conditions exist under which an IV estimate corresponding to a marginal subgroup generalizes to the patient population represented by the sample used in the analysis:¹²

1. Treatment effects are homogeneous; or
2. Treatment effects are heterogeneous but unrelated to treatment assignment.

In the first case, treatment effects are the same for everyone, making it trivial to generalize from marginal patients to all patients represented by the sample. In the second circumstance, the variables in Equation 2 affecting treatment assignment are not the same as those influencing treatment effects in Equation 1.

However, both conditions are frequently violated in clinical practice. First, if patient characteristics affect the physiological mechanisms through which treatment influences outcomes, then treatment effects are likely to be heterogeneous. Second, if the patient characteristics used in selecting treatments are unrelated to those that determine treatment effectiveness, then physicians' decisions are arbitrary. This is most likely in cases where there is greater uncertainty about treatment effectiveness.

MONTE CARLO ANALYSIS

In this section, we use Monte Carlo methods to generate data designed to resemble those collected in nonexperimental studies of medical outcomes.

Unlike methodologies used in the real world, Monte Carlo methods allow researchers to control the data-generating process and to assess how well various estimators measure the true values of model parameters. Using hypothetical data, we illustrate the generalizability of average treatment effects measured with IVs and the circumstances under which IV estimates can be used to predict the benefits of two hypothetical policy interventions that expand treatment capacity.

Monte Carlo Data

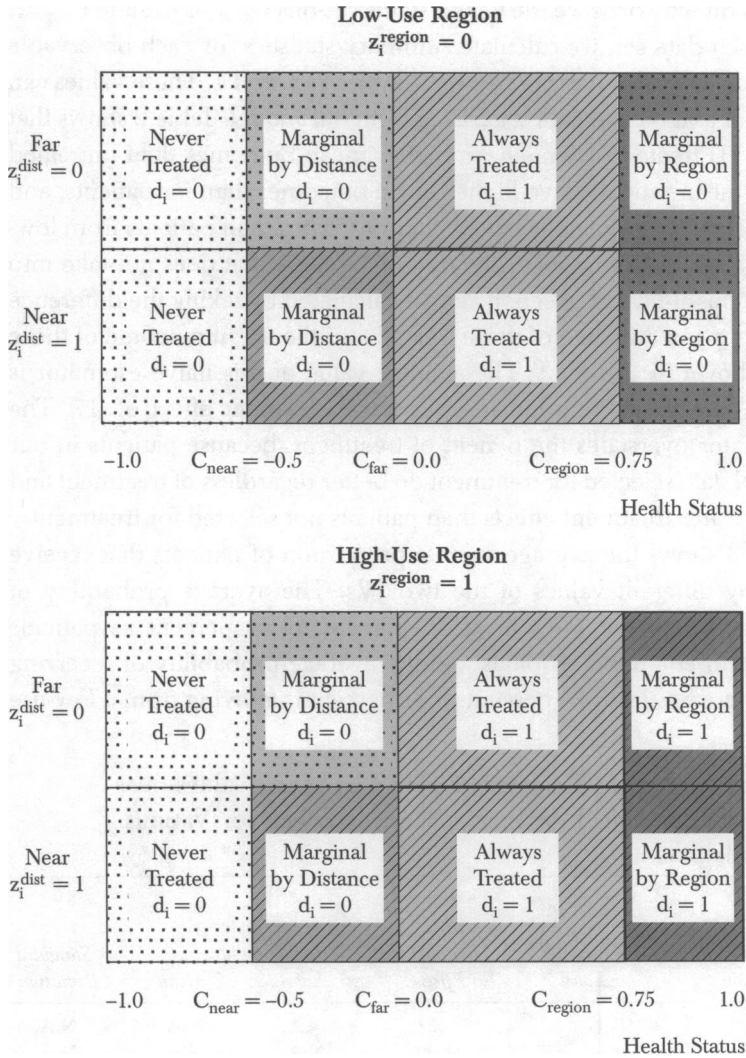
We generate our hypothetical data by assigning parameter values and distributional assumptions to the variables in the medical outcomes model described in Equations 1 and 2. (Technical details are given in the Appendix.) Our data contain two valid IVs. The first represents distance to a hospital equipped to perform treatment as described in the last section. The second represents regional differences in practice style.¹³ Figure 2 shows various subpopulations of patients when distance and regional practice style together determine treatment assignment. Again, diagonal lines represent patients who receive treatment, and dotted regions represent patients who do not receive treatment.

We imagine that physicians do not treat the healthiest patients, those with health status greater than $C_{region} = 0.75$, in the low-use region and that they do treat them in the high-use region. Regional practice style takes on two values: $z_i^{region} = 1$ if the patient lives in a high-use region and $z_i^{region} = 0$ otherwise. There are now two groups of marginal patients: (1) those whose treatment status is determined by distance, shown in the lightly shaded area of Figure 2, and (2) those whose treatment status is determined by region, shown in the darkly shaded area of Figure 2.

From this model, we draw 100 data sets of 5,000 observations each. We report true average treatment effects in the third column of Table 1 for various subpopulations corresponding to different regions of Figure 2. These values serve as benchmarks against which to compare IV estimators. Table 1 shows that the imaginary treatment is most effective for patients who always receive treatment and for those who are marginal by region. The average treatment effect for both groups is equal to 1.00. The treatment is least effective for those who are never treated. The average treatment effect for the never treated group is equal to -1.92 .

We imagine that researchers with access to a data set generated by our model can observe only treatment status, the values of the two IVs, and health outcomes that are measured with a continuous functional health status

Figure 2: Treatment Status by Health Status and Value of Two Instrumental Variables Representing Distance to a Treatment Facility and Regional Practice Style



Note: See Figure 1 for explanation of shading and patterns.

measure where the smaller numbers indicate lower function. Researchers cannot observe any aspect of health status or the process by which patients are assigned to treatments. The theory of the treatment assignment process may help researchers to divide a patient population into groups, as in Figure 2, but they cannot directly observe the health status cut-offs C_{near} , C_{far} , and C_{region} .

For each data set, we calculate summary statistics for each observable variable and report the average values in Tables 2, 3, and 4. These values can be thought of as a typical data set arising from our model. Table 2 shows that on average (1) treated patients have higher mean outcomes than untreated patients, (2) near patients have higher mean outcomes than far patients, and (3) patients in high-use regions have better outcomes than patients from low-use regions. A naive estimate of the treatment effect that does not take into account the possibility of selection can be calculated by taking the difference between the mean outcome of those treated and the mean outcome of those untreated shown in Table 2.¹⁴ The average value of this naive estimator is 1.58, far in excess of the population average treatment effect of .17. The naive estimator overstates the benefit of treatment, because patients in our hypothetical data selected for treatment do better regardless of treatment and thus have greater treatment effects than patients not selected for treatment.

Table 3 shows the average sample proportion of patients that receive treatment by different values of the two IVs. The average probability of receiving treatment increases from 22 percent to 34 percent as patients move closer to equipped hospitals and the average probability of receiving treatment increases from 25 percent to 31 percent in moving from a low-use

Table 1: True Average Treatment Effects and Mean Values over 10 Data Sets of Estimated Average Treatment Effects for Various Patient Populations

<i>Patient Population</i>	<i>Parameter Label</i>	<i>True Average Treatment Effect</i>	<i>Estimator Label</i>	<i>Mean</i>	<i>Standard Deviation*</i>
All patients	$\hat{\beta}_{pop}$	.17	N.A.	N.A.	N.A.
Always treated	$\hat{\beta}_{always}$	1.00	N.A.	N.A.	N.A.
Never treated	$\hat{\beta}_{never}$	-1.92	N.A.	N.A.	N.A.
Marginal by distance	$\hat{\beta}_{marginal}^{dist}$	.58	$\hat{\beta}_{IV}^{dist}$	.57	.14
Marginal by region	$\hat{\beta}_{marginal}^{region}$	1.00	$\hat{\beta}_{IV}^{region}$	.99	.28

*All average asymptotic standard errors were within 3 percent of the empirical standard deviations.

Table 2: Average Value over 100 Data Sets of Mean Outcomes by Treatment Status, Distance, and Regional Practice Style

<i>Observable Variable</i>	<i>Average Mean Outcome</i>	<i>Standard Deviation</i>
<i>Treatment Status</i>		
Treated	1.20	.02
Untreated	-0.38	.02
<i>Distance</i>		
Near	.58	.03
Far	.44	.03
<i>Regional Practice Style</i>		
High use	.57	.03
Low use	.45	.03

region to a high-use region. Thus, distance and regional practice style both influence the probability of being treated, satisfying one of the conditions for a valid IV.

Table 4 shows the average probability of receiving treatment conditional on different values of the two IVs. These values are needed to calculate the proportion of patients made marginal by each of the two instruments that form the denominator of the IV estimator shown in Equation 9. The average proportion of patients made marginal by distance is 69 percent – 44 percent = 25 percent. Similarly, the proportion of patients made marginal by regional practice style is 63 percent – 50 percent = 13 percent.

IV Estimates of Average Treatment Effects for the Marginal Subgroup

Next, we compare two alternative IV estimators to the true average treatment effects shown in Table 1. For each of the 100 data sets, we use distance and regional practice style to calculate IV estimates of average treatment

Table 3: Average Probability of Receiving Treatment by Distance and Regional Practice Style over 100 Data Sets

<i>Treatment Status</i>	<i>Total</i>	<i>Distance</i>		<i>Regional Practice Style</i>	
		<i>Near</i>	<i>Far</i>	<i>High Use</i>	<i>Low Use</i>
Treated	.56	.34	.22	.31	.25
Untreated	.44	.16	.28	.19	.25

Table 4: Average Probability of Receiving Treatment, Conditional on Distance and Regional Practice Style over 100 Data Sets

<i>Treatment Status</i>	<i>Distance</i>		<i>Regional Practice Style</i>	
	<i>Near</i>	<i>Far</i>	<i>High Use</i>	<i>Low Use</i>
Treated	.69	.44	.63	.50
Untreated	.31	.56	.37	.50

effects; we report the average values in the fifth column of Table 1.¹⁵ The true average treatment effect for the subgroup made marginal by distance equals .58 and is well estimated by the distance IV estimator whose average value equals .57. Similarly, the true average treatment effect for the subgroup made marginal by region equals 1.00 and is well estimated by the region IV estimator whose average value equals .99. Even though both IVs measure the average treatment effect among marginal patients, each corresponds to a different subpopulation of patients. This difference comes about because the two different IVs randomize different groups with different health status characteristics. Figure 2 shows that the health status of the group made marginal by region is higher than the health status of the group made marginal by distance. Finally, note that both IV estimators overstate the average effect of the treatment in the entire hypothetical patient population.

IV Estimates of Population Average Treatment Effects

In order for average treatment effects estimated with IV to generalize to the entire patient population represented by the sample used in an analysis, treatment effectiveness cannot depend on the same patient characteristics that affect treatment status. To illustrate this point, we generated data from a second model in which treatment effectiveness, β in Equation 1, no longer depends on health status. Since health status continues to affect both outcome in the absence of treatment ($g(h_i)$ in Equation 1) and treatment assignment ($f(h_i)$ in Equation 2), the sample selection problem persists. In the new model, treatment effectiveness continues to be heterogeneous, but in a way that is unrelated to h_i . One can think of this as "spirit," where treatment is more effective among those in "good spirits" than among those in "bad spirits." However, spirit does not affect treatment assignment. (Details are given in the Appendix.) Because treatment effectiveness depends on spirit, not on health status, we expect the marginal group in each of our 100 data sets to contain the same mixture of patients in good spirits and in bad spirits as the

entire patient population. Thus, the effect of treatment estimated with an IV generalizes without difficulty to the population as a whole.

Table 5 shows the results of the Monte Carlo simulation with the new “spirit” model. Unlike the case where health status affected both treatment assignment and treatment effectiveness, the average treatment effects for the two marginal subgroups are now the same as the average treatment effect for the entire population. Consequently, both IV estimates well approximate the average treatment effect over the whole population.

Predicting the Effects of Policy Interventions and Other Changes in Treatment Patterns

Medical outcomes research can inform a range of policy questions. Should a particular form of treatment be used? Who should receive treatment? Should certain facilities be built? This section shows that IV estimates are useful for predicting the impact on outcomes of policy interventions that alter practice patterns only when these changes affect the treatment status of roughly the same subgroup whose treatment status is determined by the value of the IV.

We illustrate this point by simulating the effect of two types of policy interventions. Both interventions increase the proportion of patients who receive treatment, but each affects a different subgroup of patients. Each can be interpreted as an expansion of the capacity of hospitals to perform the treatment of interest. We can think of the first intervention as occurring in a rural setting, while the second can be thought of as taking place in an urban setting.

It is reasonable to think that when capacity is expanded, physicians will begin to treat patients about whom they are uncertain of the effectiveness of treatment. In principle, the newly treated groups could be well represented

Table 5: True and Estimated Average Treatment Effects When Effectiveness Is Unrelated to Treatment Assignment Process

<i>Patient Population</i>	<i>Parameter Label</i>	<i>True Average Treatment Effect</i>	<i>Estimator</i>	<i>Mean</i>	<i>Standard Deviation*</i>
All patients	$\bar{\beta}_{pop}$	.17	N.A.	N.A.	N.A.
Marginal by distance	$\bar{\beta}_{marginal}^{dist}$	.17	$\hat{\beta}_{IV}^{dist}$	.16	.17
Marginal by region	$\bar{\beta}_{marginal}^{region}$	.17	$\hat{\beta}_{IV}^{region}$	.23	.38

*Both average asymptotic standard errors were within 3 percent of the empirical standard deviations.

by either of our two marginal subgroups, because both groups consist of patients whose treatment status is largely determined by factors other than health status—in our example, distance and available resources.

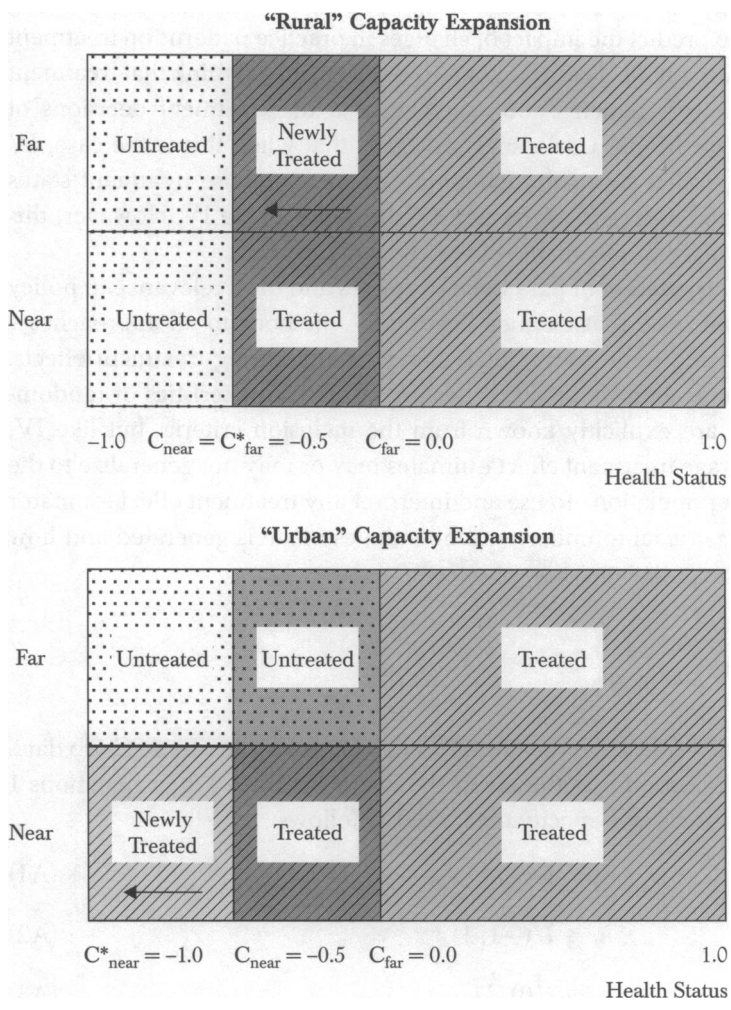
In conducting the two simulations, we return to the medical outcomes model described earlier in this section where health status influences both treatment effectiveness and treatment assignment. First, we simulate the effect of an expansion of capacity in a rural area by lowering the health status cutoff for those who live far from equipped hospitals from C_{far} to C_{far}^* , while leaving unchanged the treatment status of those who live nearby (as illustrated in the top of Figure 3). The formerly marginal group of patients who were not treated prior to the intervention now become treated. The average value of the treatment effect of the newly treated group $\hat{\beta}_{rural}$ equals .58. Because the newly treated subgroup is the same as the subgroup whose treatment status is determined by distance, $\hat{\beta}_{IV}^{dist}$ correctly predicts the benefits of the capacity expansion: the average value of $\hat{\beta}_{IV}^{dist} = .57$, shown in Table 1, is almost identical to that of $\hat{\beta}_{rural}$. Note that $\hat{\beta}_{IV}^{region} = .99$ overstates the benefits of the expansion. This is because treatment is much more effective for the regionally marginal group than it is for the group made marginal by distance.

Next, we simulate the effect of a capacity expansion in an urban area by lowering the health status cutoff for those who live near equipped hospitals from C_{near} to C_{near}^* , while leaving unchanged the treatment status of those who live far from them (as illustrated in the bottom of Figure 3). The formerly never treated near patients become treated. The average value of the treatment effect of the newly treated group $\hat{\beta}_{urban}$ equals -1.92 . Because the health status of newly treated patients is much lower than the health status of either of the two marginal groups, both $\hat{\beta}_{IV}^{dist}$ and $\hat{\beta}_{IV}^{region}$ overstate the benefits of the capacity expansion.

In real life, researchers cannot identify the subgroup of candidates whose treatment status is determined by an instrument. Our simulations highlight the importance in using IV methods to inform policy decisions because they are able to justify on theoretical grounds the fact that IV affects the same population that will be affected by the policy intervention.

However, in certain circumstances the policy intervention changes treatment status in the same way as an available IV. Consequently, the IV estimate directly informs the policy question without raising issues of generalizability. For example, malpractice law may vary across states in a random fashion, enabling the use of state variation in malpractice law as an IV. The resulting IV estimate directly predicts the effect of the law on the policy-relevant subgroup.

Figure 3: Effect of Expanded Capacity on Treatment Status by Health Status and Value of Instrumental Variable Representing Distance to Hospital



Note: See Figure 1 for explanation of shading and patterns.

DISCUSSION

Researchers using IV methods to measure average treatment effects should be aware of generalizability issues in interpreting their results and in using their results to predict the impact of changes in practice patterns on treatment effectiveness. In medical care contexts, it is reasonable to think that treatment effects vary across patients and are related to the treatment decisions of physicians. Our Monte Carlo analysis shows that when this is the case, IV estimates generalize to a subpopulation of patients whose treatment status was determined in a random fashion by the value of the IV. However, this subpopulation is not observable in readily available data.

The interpretation of parameter estimates and their relevance to policy questions is an issue that extends beyond IV methods to all approaches—experimental and nonexperimental—aimed at measuring treatment effects. In experimental settings, unlike IV, many of the characteristics of randomized subjects are explicitly known from the inclusion criteria, but like IV, resulting average treatment effect estimates may or may not generalize to the larger patient population. To use and interpret any treatment effect estimator correctly, it is crucial to understand how the estimate is generated and how it relates to the policy question of interest.

APPENDIX

In this section, we describe the model used to generate the Monte Carlo data. Our model corresponds to the medical outcomes framework in Equations 1 and 2 of this article. We specify the model as follows:

$$y_i = \beta(h_i)d_i + h_i + \varepsilon_i \quad (\text{A1})$$

$$h_i \sim U(-1, 1) \quad (\text{A2})$$

$$\varepsilon_i \sim N(0, 1) \quad (\text{A3})$$

$$\beta(h) = \begin{cases} 1 - 5h^2 & \text{if } h \leq 0 \\ 1 & \text{if } h > 0 \end{cases} \quad (\text{A4})$$

This functional form is based on the notion that very ill patients are fragile and that treatment worsens their outcome. As patients' health improves, treatment becomes more effective until some maximum benefit level is reached and the benefit remains constant.

$$z_i^{dist} = \begin{cases} 1 & \text{if } \phi_i \leq 0 \text{ near} \\ 0 & \text{if } \phi_i > 0 \text{ far} \end{cases} \quad (\text{A5})$$

$$\phi_i \sim U(-1, 1) \quad (\text{A6})$$

$$z_i^{region} = \begin{cases} 1 & \text{if } x_i \leq 0 \text{ low use} \\ 0 & \text{if } x_i > 0 \text{ high use} \end{cases} \quad (\text{A7})$$

$$x_i \sim U(-1, 1) \quad (\text{A8})$$

$$d_i = \begin{cases} 0 & \text{if } h_i \leq c_1(z_i^{dist}, z_i^{region}) \\ 1 & \text{if } c_1(z_i^{dist}, z_i^{region}) < h_i \leq c_2(z_i^{dist}, z_i^{region}) \end{cases} \quad \text{or } h_i > c_2(z_i^{dist}, z_i^{region}) \quad (\text{A9})$$

$$c_1(1, z^{region}) = C_{near} = -0.5, \quad c_1(0, z^{region}) = C_{far} = 0.0, \\ c_2(z^{dist}, 1) = C_{region} = .75, \quad c_2(z^{dist}, 0) = 1.0 \quad (\text{A10})$$

When “spirit” determines treatment effectiveness, the model is identical to the one above with the following exceptions:

(1) There is a new variable spirit,

$$s_i \sim U(-1, 1) \quad (\text{A11})$$

(2) Equation A4 becomes

$$\beta(s_i) = \begin{cases} 1 - 5s_i^2 & \text{if } s \leq 0 \\ 1 & \text{if } s > 0 \end{cases} \quad (\text{A12})$$

(3) Equation A1 becomes

$$y_i = \beta(s_i)d_i + .25s_i + h_i + \varepsilon_i \quad (\text{A13})$$

ACKNOWLEDGMENTS

The authors gratefully acknowledge the comments and suggestions of Bryan Dowd, Roger Feldman, Robert Kane, Lawrence Katz, Eric Lawrence, John LeFante, Joseph Newhouse, and two anonymous referees.

NOTES

1. Two-stage least squares is one form of instrumental variables estimation.
2. Strictly, weaker conditions than strict randomness are needed for a valid IV. This distinction is discussed briefly later in the article.

3. The authors use a variable called "differential" distance, measuring the difference between the nearest hospital equipped to perform cardiac catheterization and the nearest hospital without such capabilities. However, this distinction is not important for the purposes of this article.
4. The authors thank an anonymous reviewer for pointing out that the IV used by McClellan, Newhouse, and McNeil (1994) is appropriate for measuring the impact of access to intensive treatment, not intensive treatment per se.
5. Angrist, Imbens, and Rubin (1996) derive these four necessary conditions mathematically.
6. The plausibility of the behavioral assumptions required for valid IVs has been the subject of ongoing debate in the program evaluation literature (see Keane and Wolpin 1997; Moffitt 1996).
7. Researchers must also ensure that large sample statistics apply (Bound, Jaeger, and Baker 1995; Staiger and Stock 1997).
8. There are weak tests of the exclusion restriction formulated under the assumption of homogenous treatment effects (see Kennedy 1992, Section 10.2, General Notes). This test requires more than one valid IV. However, the substantive interpretation of the model under the assumption of heterogeneous treatment effects with more than one valid IV is problematic (Heckman 1997), and thus the meaning of the test is unclear.
9. The monotonicity assumption can be thought of as the price that one must pay for relaxing the strict assumption that treatment effect heterogeneity is unrelated to the treatment assignment process. Heckman (1997) shows this formally.
10. Note that under the assumptions required for a valid IV, it is possible to determine the relative size of the marginal subgroup. However, this does *not* imply that the marginal subgroup is observable.
11. For ease of exposition, we use the nonparametric form of the IV estimator presented by Angrist (1990) and Wald (1940) that uses both a dichotomous instrument and a dichotomous treatment status variable. There exist other formulations of IV estimators that are appropriate for continuous data that require less general assumptions about the relationship between health outcomes and treatment effects. Two-stage least squares (2SLS) is one form of IV estimator familiar to many health services researchers. See Kennedy (1992) for a discussion of the 2SLS estimator. The intuition about the generalizability of treatment effect estimates in the 2SLS case remains the same as in the nonparametric case. Imbens and Angrist (1994) show formally the relationship between the nonparametric and parametric estimators.
12. Heckman (1997) discusses these conditions.
13. As noted earlier, there is reason to believe that region may not be a valid instrument in real world data.
14. With our simple dichotomous variables, the naïve estimator is equivalent to the ordinary least squares estimator.
15. IV estimates can be calculated using the information in Tables 2 through 4 and Equation 7. However, the estimators calculated from the average data set are

slightly different from the average of the 100 estimators because the IV estimates are nonlinear functions of treatment probabilities.

REFERENCES

- Andersen, C. 1994. "Measuring What Works in Health Care." *Science* 263 (5150): 1080.
- Angrist, J. D. 1995. "Introduction to the JBES Symposium on Program and Policy Evaluation." *Journal of Economics and Business Statistics* 13 (2): 133–36.
- . D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80 (3): 313–35.
- Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.
- Bound, J., D. A. Jaeger, and R. M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90 (430): 443–50.
- Currie, J., and J. Gruber. 1996. "Health Insurance Eligibility, Utilization of Medical Care and Child Health." *Quarterly Journal of Economics* 111 (2): 431–66.
- Davis, K. 1995. Letter to the Editor. *Journal of the American Medical Association* 273 (17): 1332.
- Gould, K. L. 1994. "Invasive Procedures in Acute Myocardial Infarction: Are They Beneficial?" *Journal of the American Medical Association* 272 (11): 891–93.
- Gross, J. S., and J. R. Shua-Haim. 1995. Letter to the Editor. *Journal of the American Medical Association* 273 (17): 1332.
- Guadagnoli, E., P. J. Hauptmann, J. Z. Ayanian, C. L. Pashos, B. J. McNeil, and P. D. Cleary. 1995. "Variation in the Use of Cardiac Procedures After Acute Myocardial Infarction." *The New England Journal of Medicine* 333 (9): 589–90.
- Heckman, J. J. 1997. "Instrumental Variables: A Study of the Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources* 32 (3): 441–62.
- Heckman, J. J., and J. A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9 (2): 85–110.
- Holland, P. W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.
- Imbens, G. W., and J. D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.
- Institute of Medicine. 1985. *Assessing Medical Technologies*. Washington, DC: National Academy Press.
- Keane, M. P., and K. Wolpin. 1997. "Introduction to the JBES Special Issue on Structural Estimation in Applied Microeconomics." *Journal of Business and Economic Statistics* 15 (2): 111–14.

- Kennedy, P. 1992. *A Guide to Econometrics*. Cambridge, MA: MIT Press.
- Kessler, D., and M. McClellan. 1996. "Do Doctors Practice Defensive Medicine?" *Quarterly Journal of Economics* 111 (2): 353-90.
- Kuller, L. H., and K. Detre. 1995. Letter to the Editor. *Journal of the American Medical Association* 273 (17): 1331-32.
- Levitt, S. D. 1996. "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Legislation." *Quarterly Journal of Economics* 111 (2): 319-52.
- McClellan, M. B., J. P. Newhouse, and B. J. McNeil. 1994. "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality?" *Journal of the American Medical Association* 272 (11): 859-66.
- Meyer, B. D. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business and Economic Statistics* 13 (2): 151-61.
- Moffitt, R. 1996. "Comment on Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 462-64.
- Quality Standards Subcommittee of the American Academy of Neurology. 1996. "Practice Advisory: Thrombolytic Therapy for Acute Ischemic Stroke—Summary Statement." *Neurology* 47 (3): 835-39.
- Robins, J. M., and S. Greenland. 1996. "Comment on Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 456-61.
- Rubin, D. B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 7 (1): 34-58.
- . 1974. "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies." *Journal of Educational Psychology* 66 (5): 688-701.
- Staiger, D., and J. H. Stock. 1997. "Instrumental Variables with Weak Instruments." *Econometrica* 65 (3): 557-86.
- von Kummer, R., K. L. Allen, R. Holle, L. Buzzao, S. Bastianello, C. Manelfe, E. Bluhmki, P. Ringleb, D. H. Meier, and W. Hacke. 1997. "Acute Stroke: Usefulness of Early CT Findings Before Thrombolytic Therapy." *Radiology* 205 (2): 327-33.
- Wald, A. 1940. "The Fitting of Straight Lines if Both Variables Are Subject to Error." *Annals of Mathematical Statistics* 11 (September): 284-300.
- Wennberg, J. E. 1987. "Illness Rates Do Not Explain Hospitalization Rates." *Medical Care* 25 (4): 354.
- Wennberg, J. E., and A. Gittelsohn. 1982. "Variations in Medical Care Among Small Areas." *Scientific American* 246 (4): 120.