HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 11/26

Exploring comparative effect heterogeneity with instrumental variables: prehospital intubation and mortality

Heather Evans
Anirban Basu

August 2011

york.ac.uk/res/herc/hedgwp

Exploring comparative effect heterogeneity with instrumental variables: prehospital intubation and mortality

Heather Evans, MD, MS[1]
Anirban Basu, PhD[2,*]

[1] Department of Surgery, University of Washington, Seattle WA

[2] Department of Health Services and Pharmacy, University of Washington, Seattle WA; National Bureau of Economic Research, Massachusetts.

August 8, 2011

Acknowledgement

We are grateful to Ellen MacKenzie, Gregory Jurkovich, and Frederick Rivara for providing access to the NSCOT data and to Jin Wang for her assistance in data preparation and analysis. We also thank Jim Heckman and Philipp Eisenhauer for suggestions about computations. All errors are ours. Dr. Basu acknowledges financial support from National Cancer Institute research grants RC4 CA155809 (PI Basu) and R01 CA155329 (PI Basu). Dr. Evans acknowledges financial support from an Agency for Health Care Quality and Research career development award K12 HS019482-01 (PI Sullivan).

* Corresponding author
Anirban Basu Ph.D.
University of Washington, Seattle
1959 NE Pacific St
Box 357660
Seattle, WA, 98195-7660
Tel: (206) 616-2986 | Fax: (206) 543-3964
basua@uw.edu

Abstract

We highlight the role of local instrumental variable (LIV) methods in exploring treatment effect heterogeneity using an empirical example of evaluating the use versus non-use of prehospital intubation (PHI) in patients with traumatic injury on inpatient mortality. We find evidence that the effect of PHI on inpatient mortality varies over levels of unobserved confounders giving rise to a phenomenon known as *essential* heterogeneity. Under essential heterogeneity, the traditional instrumental variable (IV) method, when using a continuous IV, estimates an effect that is an arbitrary weighted average of the casual effects for marginal groups of patients whose PHI receipt are directly influenced by the IV levels. Instead, the LIV methods estimate the distribution of treatment effects for every margin that is identified by data and allow for predictable aggregation to recover estimates for meaningful treatment effect parameters such as the Average Treatment Effect (ATE) and the Effect on the Treated (TT). LIV methods also allow exploring heterogeneity in treatment effects over levels of observed confounders. In the PHI analysis, we estimate an ATE of 0.074 (se=0.02, $p<0.001$) and a TT of -0.079 (se=0.09, $p=0.38$). We find strong evidence of positive self-selection in practice based on observed and unobserved characteristics, whereby patients who were most likely to be harmed by PHI were also less likely to receive PHI. However, the degree of positive self-selection mitigates in regions with higher rates of PHI use. We also explore factors associated with the prediction of significant harm by PHI. We provide clinical interpretation of results and discuss the importance of these methods in the context of comparative effectiveness research.

Key words: Instrumental variables; local IV methods; heterogeneity; prehospital intubation; mortality.

Introduction

With enriched resources available for conducting comparative effectiveness research (CER) in the United States and the continuous development of more comprehensive observational databases based on electronic medical records, statistical and econometric methods for estimating treatment effects are in great demand. Treatment effects are the primary parameters of interest in comparative effectiveness research, albeit some effects are more useful for linking results to decision making within health care than others. Since the goal of CER is to help make better decisions within health care [1], two fundamental requirements arise in the development of such methods: 1) the treatment effect parameters reflect the needs of the specific decision maker in question, and 2) the estimates for these parameters carry a causal interpretation.

In this paper, we explore the comparative effectiveness of prehospital intubation (PHI) compared to no PHI on inpatient mortality following traumatic injury among patients who reached the emergency department alive. The safety and efficacy of PHI in trauma patients is controversial; there are currently no definitive guidelines as to when it is better to perform intubation at the scene by first responders or to defer definitive airway management until after arrival at a hospital to facilitate the most rapid conveyance to advanced medical care. A registry-based retrospective cohort study demonstrated increased risk of death and long-term disability in patients with severe traumatic brain injury (TBI) who were intubated prior to arrival to the hospital, compared to those intubated in the emergency department (EDI) [2]. In contrast, the largest single-center study, comparing PHI to EDI, demonstrated no significant difference in rates of ventilator associated pneumonia or death [3]. It was suggested that the results, contrary to those of prior retrospective cohort studies demonstrating a higher incidence of pneumonia in prehospital-intubated patients, may have been influenced by a well-established, standardized paramedic rapid sequence intubation training program with continuous quality improvement measures.

The problem with these studies, as is widely recognized as the fundamental challenge of estimating treatment effects from observational data, is selection bias. Selection bias (i.e., confounding by indication) arises when factors that can influence the treatment choice, such as patient health, resource availability and provider skills, also influence 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 was widely adopted. Use then plummeted when these studies were eventually disproven by a large randomized trial [4]. It has been shown

subsequently that the reason for the discrepant results was that the observational studies failed to consider certain confounders such as socioeconomic status [5] or failed to distinguish initiation of therapy from prevalence of therapy [6]. The significance of overcoming the limitations 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. Unfortunately, a study with a randomized design, which can ideally overcome these selection biases, may be logistically and/or ethically challenged in the context of many crucial clinical questions, including the case of when to perform intubation in patients with traumatic injury. Even in the case of the only RCT performed to date in a traumatic brain injury population, blinding as to the treatment group amongst providers was not possible (as the intubation could not be concealed), potentially affecting the results of the study [7].

In order to address selection bias, in line with what a pragmatic randomized design should have accomplished, we will focus our attention to the use of instrumental variables (IV). In what follows, we highlight the role of traditional and newer IV methods in comparative effectiveness research, substantiated with evaluating the effectiveness of prehospital intubation using these methods.

Overview of Instrumental Variable Methods and Interpretation of Results

Instrumental variable (IV) methods have been a cornerstone method for observational studies in the economics literature and its origins date back to the 1920s [8]. In the last couple of decades, these methods have gained popularity in the medical literature on the evaluation of alternative medical treatments [9,10,11,12,13], 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 [14,15,16]. 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 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.

Non-essential heterogeneity

The field of CER is also devoted to estimating heterogeneity in treatment effects [17]. In the presence of treatment effect heterogeneity, however, results from traditional IV approaches may suffer from lack of interpretability. 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 possess that can enable them to anticipate these effects and select into treatment that would potentially give them the largest benefit. In other words, unobserved confounders are assumed not to be moderators of treatment effects (this situation is denoted as non-essential heterogeneity [20]).

An underlying data generating mechanism for non-essential heterogeneity is illustrated with a stylized example of potential outcomes in Figures 1(a) and 1(b). In these figures, the X-axis represents levels of an unobserved confounder, while the Y-axis represents the potential outcomes. The line connecting the '+'s represents the schedule of potential outcomes in the population had every patient received the control treatment. As constructed, people respond differently to the same treatment. This is called response heterogeneity. The line connecting circles represent the schedule of potential outcomes for the same people had they received the new treatment. Although there is also response heterogeneity from the new treatment, the differential responses across alternative treatments (denoted by the grey bars in Figure 1(a)) are held constant across people. That is, treatment-effect heterogeneity is constant across levels of the unobserved confounders. Technically, treatment-effect heterogeneity is denoted as non-essential only when it is statistically independent of response heterogeneity.[1] Figure 1(b), illustrates how IVs produce interpretable results in this situation. In practice, we do not observe the potential outcomes under both treatments for each patient. Rather, we observe outcomes for a self-selected group of patients receiving each treatment. An IV helps match the unobserved level of confounding. Therefore, an IV compares outcomes for a treated and an untreated group of patients, whose treatment choices are driven by the levels of the IV, and

[1] Note that for non-essential heterogeneity, there should be full independence between response and treatment effect heterogeneity, and not just mean independence.

hence the levels of their unobserved confounders are held fixed at some arbitrary value defined by the specific instrument used [18]. For example, an analyst using distance to treatment facilities as an instrument would, in effect, hold the levels of unobserved confounding fixed at a specific level, which may be different that the level held fixed by another analyst using physician preferences as instruments. However, with non-essential heterogeneity, the level of unobserved confounder at which an instrument is 'acting' is inconsequential, since each IV will estimate the constant treatment effect, which is also the average treatment effect in the population. The interpretation of IV estimates in such a situation is straightforward.

Essential Heterogeneity

When unobserved confounders moderate treatment effect in a systematic fashion, treatment effect heterogeneity depends on response heterogeneity, and becomes known as *essential heterogeneity*. This is illustrated in Figure 1(c). Since the marginal patients identified by an IV are entirely dependent on the specific instrument being used and how this instrument affects treatment choices [15, 16], the use of different instruments by different analysts 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 [19], 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) or arbitrarily weighted averages of LATEs. This estimate is often substantially different from mean treatment effect concepts such as the Average Treatment Effect (ATE).

A new genre of IV methods originally developed by Heckman and colleagues [20,21,22] directly addresses these limitations of traditional methods. Known as local instrumental variable (LIV) approaches, these methods can relax assumptions, allow unobserved characteristics of patients that influence treatment choices to also be moderators of treatment effects, and recover the full distribution of treatments effects across all possible margins of patients choices, not just the one directly influenced by an IV, by explicitly developing 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 who choose treatment and some who 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 (Figure 1(d)). 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). 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 [23]. 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. For CER, there is strong theoretical reasoning for why a treatment effect that is averaged over all patients in a population, i.e. the Average Treatment Effect (ATE), can mislead patient or physician decision making, ultimately affecting welfare in this population [24,25,26]. Therefore, more nuanced subgroup specific effects, represented by conditional MTEs, are often more useful.

An LIV method can confirm if the assumption of non-essential heterogeneity is valid. It also provides a seamless approach to explore treatment effect heterogeneity across observed confounders. Recently, Basu et al. [27,28] applied these methods to estimate ATE, TT and MTEs of breast cancer treatments on costs and mortality. A detailed description of the theory and methods on LIV approaches as it relates to CER can also be found in these works.

In this paper, we apply traditional IV and LIV approaches to estimates interpretable treatment effects and also to explore heterogeneity in effects of the use of prehospital intubation compared to emergency department intubation on inpatient mortality following traumatic injury.

Clinical Context of Prehospital Intubation and Mortality

Patients who sustain injuries are susceptible to aspiration and loss of airway due to decreased level of consciousness, whether due to direct head trauma or other severe injury resulting in shock. Prevention of secondary brain injury by avoidance of hypoxia and hypotension is a primary goal in the initial treatment of head trauma, and early intubation has been advocated to facilitate improved oxygenation [29]. However, in 2007, an expert panel concluded that there was not sufficient data to promote the standard practice of PHI for patients with traumatic brain

injury [30]. These recommendations were based on the available data, including multiple single-center retrospective analyses with opposing conclusions as to the harm or benefit of PHI. For example, a review of severely head injured patients in a statewide trauma registry revealed a 4-fold adjusted risk of death for PHI versus emergency department intubation [2]. Although the authors contend they had employed the best available risk-adjustment methods including propensity scoring, unobserved differences in the two groups may have influenced the decision to intubate as well as the observed outcomes. A single prospective observational study of prehospital rapid sequence intubation in 209 patients matched to historical non-intubated controls concluded PHI was associated with increased mortality, possibly due to inadvertent hyperventilation, transient hypoxia, or longer time at scene prior to transport [31]. Accordingly, the panel suggested that all of the studies considered failed to account for potentially important confounders, among the many methodological issues that hampered their ability to definitively recommend for or against PHI [30]. It concluded that a randomized control trial was one of the main goals for future investigation.

While not impossible to accomplish, because of the time-critical factors and the subject's inability to consent to randomized treatment, out-of-hospital clinical trials necessitate waiver of consent and are among the most highly regulated studies. This level of scrutiny may be, at least in part, the reason it took four years to enroll only 312 patients, despite 1045 screened, in the only randomized controlled trial of prehospital intubation in adult patients with severe head injury reported to date [7]. It is widely recognized that RCTs may not allow for real-world conduct of treatments in these settings, and may significantly vary from what is possible or practical to perform in practice. In the study by Bernard et al [7], patients who did not receive rapid sequence intubation medications were excluded. Even in employing the most-highly trained paramedics already certified to perform advanced airway management, the study required an additional 16 hours of airway management training for study participants to learn rapid sequence intubation techniques. It has been previously reported that in addition to training and maintenance of skills, ongoing quality assurance is required for successful administration of a PHI program [32]. Whether prehospital trauma provider systems adopt such measures will undoubtedly be impacted by resource availability and the perceived potential benefit to the patients served.

The effect of PHI may vary over numerous factors, many of which cannot be measured due to the current limitations of prehospital data collection and reporting. Variability in the success of PHI may be influenced by patient factors (both pre-existing and injury-related conditions), provider skill level and/or experience in decision making, resource availability, and

environmental constraints at the scene of injury. We set out to explore heterogeneity in comparative effects across both the observed and unobserved factors that also may have played a role in the choice of intubation in practice. We use data from the National Study on Costs and Outcomes in Trauma (NSCOT) for this purpose. The planned analyses of this data examined variation in care delivery between level I trauma centers and non-trauma centers, determined the extent to which differences in care correlate with patient outcome (including major functional outcomes at 3 and 12 months after injury), and estimated acute and 1-year costs, describing the relationship between costs and effectiveness for trauma and non-trauma centers [47,33,34,35,36,37,38,39]. As a rich source of prospectively collected observational injury-related data, the NSCOT has also been used for a number of secondary analyses, beyond the originally intended scope of the study [40,41]. In 2007, Bulger and colleagues used the NSCOT data to demonstrate significant variation in the conduct of out-of-hospital treatments [42]. The authors observed variation in the rates of endotracheal intubation across regions, ranging from 5% to 48% of all patients treated, and noted that the variation persisted after stratification by severity of injury, suggesting that this wide range in practice was not dependent upon patient injury heterogeneity observed across regions. To date, an IV analysis has never been used on this data. Moreover, the comparative effects of PHI versus no PHI on inpatient mortality have rarely been studied in trauma patients, with most retrospective analyses focusing on the comparison between patients intubated in the field prior to arrival at the hospital to those intubated in the emergency department.

Theory of Local Instrumental Variable (LIV) Approach

The theory of local instrumental variable (LIV) approach starts with a formal model for choices. Let the net (latent) utility for treatment,[2] Λ, based on which choices are determined,[3] is given as

[2] 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 vary over all factors. A factor cannot affect treatment choice unless it affects this latent utility.

An alternative formulation of the choice model that is used in Heckman et al. [22] is

$$\Lambda = \mu_{\Lambda}(Z,X) - U_{\Lambda} \qquad E(U_{\Lambda})=0 \qquad D = 1(\Lambda > 0),$$

where the subtraction of the error term U_{Λ} makes P(Z) enter as an upper limit of the CDF for U_{Λ}. However, most traditional econometric software packages fit (1) and not the model with a negative error term. This leads to a slight differences in the way the values of U_{Λ} are evaluated. We follow the model in (1) that is also implanted in Stata®.

[3] Decision maker in the intubation context is likely the paramedic attending the trauma patient.

$$\Lambda = \mu_\Lambda(X,Z) + U_\Lambda \qquad \mathrm{E}(U_\Lambda) = 0 \qquad D = I(\Lambda > 0), \tag{1}$$

where X represents a vector of observed confounders and Z represent a vector of instrumental variables. $U_\Lambda = \Lambda - \mu_\Lambda(X,Z)$ has expectation of zero while $I(.)$ is an indicator function representing treatment choice D. Equation (1) expresses the typical random utility framework for discrete choices in econometrics [43,44]. Following this framework, one can write $D = I(\Lambda > 0) = 1(U_\Lambda > -\mu_\Lambda(z,x)) \Leftrightarrow$, $1(F_{U_\Lambda}(U_\Lambda) > F_{U_\Lambda}(-\mu_\Lambda(z,x))) \Leftrightarrow 1(F_{U_\Lambda}(U_\Lambda) > 1 - P(z,x))$ where $P(z,x) = F_{U_\Lambda}(\mu_\Lambda(z,x))$ and $F_{U_\Lambda}(U_\Lambda) = U_D \sim Uniform(0,1)$ by construction. The formulation in (1) decomposes factors that determine choice of treatment into the observed and unobserved components (again, by the analyst). The additive separability of (1) in terms of observables and unobservables plays a crucial role in the justification of instrumental variable methods [20,22]. Hereon, we denote $S(z,x) = 1 - P(z,x)$. At the indifference point, $S(z,x)$ must balance off $U_D = u_D$. That is, in order to bring a decision making agent to indifference, the numerical values of $S(z,x)$ and u_D must be the same; therefore, a high $P(z,x)$ is needed to compensate a low $U_D = u_D$. Since high $P(z,x)$ indicates higher likelihood of getting treatment based on observed characteristics, a low u_D must indicate a lower likelihood of getting treatment based on unobserved characteristics.

Consider for simplicity the single instrument case, i.e. Z is a scalar rather than a vector of instruments. Given model (1) and the assumed independence of Z and U_Λ, changing Z externally from U_Λ, shifts all people in the same direction (towards or against $D = 1$). This produces “monotonicity" in the sense of Imbens and Angrist [18].

If Y_1 and Y_0 represent the potential outcomes for a patient with treatment and control respectively, the treatment effect for that patient is denoted as $\Delta = Y_1 - Y_0$. A Marginal Treatment Effect (MTE) [45,16,20] 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 as:

$$\begin{aligned}
\mathrm{MTE}(x,z) &= E(\Delta \mid X = x, Z = z, \Lambda = 0) = E(\Delta \mid X = x, U_\Lambda = -\mu_\Lambda(z,x)) \\
&= \mu_1(x) - \mu_0(x) + E(U_1 - U_0 \mid U_\Lambda = -\mu_\Lambda(z,x)) \\
&= \mu_1(x) - \mu_0(x) + E(U_1 - U_0 \mid U_D = S(z,x)),
\end{aligned} \tag{2}$$

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. For example, a local

average treatment effect (LATE) [15] 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 [20,22].

An additional feature of MTE is that all mean treatment effects parameters, including the Average Treatment Effect (ATE), Effect on the Treated (TT), and the traditional IV effect, can be calculated from weighted averages of MTE. These weights can be obtained from the data at hand [22,27,46]. For example, the ATE is the sum of all MTE across all distinct values of U_D, weighted equally (conditional on X).

If MTEs do not vary over U_D, it provides direct evidence on the absence of essential heterogeneity. In such a situation, the conditional MTEs converge to the conditional ATE or the TTs. One can then solely focus on exploring treatment effect heterogeneity across observed confounders.

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 [20,22,46]. Specifically, the rate of change of the mean outcome with respect to $P(Z)$, where the variable $P(Z,X)$ is evaluated at particular values of $p(z,x)$ gives

$$\frac{\partial}{\partial P(z,x)} E(Y \mid Z=z, X=x)\Big|_{P(Z,X)=p(z,x)} = E((Y_1 - Y_0) \mid X=x, U_D = 1-p(z,x))$$

$$= E\{\Delta \mid X\} + \frac{\partial K(P(z,x))}{\partial P(z,x)}\Big|_{P(Z,X)=p(z,x)} = \text{MTE}(x, U_D = 1-p) \qquad (3)$$

where $E\{\Delta \mid X\}$ is the average treatment effect conditional on X and $K(P(z,x))$ is a differentiable function of $P(z,x)$. A formal derivation is given in the Appendix. Equation (3) 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, equation (3) 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, as needed, to fully capture the non-linearity of the outcome with respect to P(Z). One then computes the partial derivative of the regression estimand with respect to the propensity score to get an LIV estimand and evaluates the argument P in the LIV estimand at $P(Z,X) = p(z,x)$ in order to produce an estimate for MTE(x,U_D = 1-p(z,x)). Thus, the support of U_D that is identified in the data would be the same as the range of (1-p(x,z)). 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.

Methods

Data

The National Study on Costs and Outcomes in Trauma (NSCOT) was a prospective observational cohort study of trauma treatment at 27 Level 1 trauma centers and 124 non-trauma hospitals in 15 US regions, including both urban and suburban centers [47,48].The 15 regions were defined by one or more contiguous metropolitan statistical areas (MSA) in 14 states. Subjects enrolled were treated for moderate to severe injury at participating hospitals between July 2001 and November 2002. Among the major exclusion criteria were patients who arrived at the hospital without vital signs and those pronounced dead within 30 minutes of arrival to the hospital. Further details about the selection criteria and sampling methodology have been previously described [47,48].

For the purpose of the analysis described herein, the NSCOT data were restricted to a subset of patients whose injury was severe enough that advanced airway management was a reasonable possibility during the acute out-of-hospital resuscitation phases (ISS≥16). This was primarily based on the range of severity of illness in a review of the available evidence in a published clinical practice guideline from the time proximate to the study period [49]. Indeed, the rate of PHI in patients with ISS<16 was less than 2% in the NSCOT data. We did not use inclusion criteria according to ‘‘need’’ for endotracheal intubation because there are no validated definitions, scales, or prediction rules for this characteristic. Furthermore, while a Glasgow Coma Score (GCS) of <9 is generally used as an indication for intubation, it was not used as an inclusion criterion in the current study because the time of GCS measurement with respect to drug administration (including muscle relaxants associated with rapid sequence intubation

protocols) was not clear, suggesting that GCS might be a result of the treatment rather than an indication for it. We also restricted the study population to those patients transported directly from the scene of injury to the hospital by either air- or ground-based prehospital medical providers. Failed PHI attempts (n=101, 19% of all PHI) were included in the PHI group, even if they were subsequently intubated in the emergency department (n=80).

The available dataset allowed us to account for a variety of confounders, including patient demographics, pre-injury health insurance status, and clinical characteristics such as presence of prehospital shock, presence of severe head injury (as measured by Head Abbreviated Injury Scale score ≥3), admission to a trauma center, and injury severity score (ISS) categories. Among the potential unobserved confounders were skill level and experience of provider performing intubation [50,51], scene characteristics that might make intubation particularly difficult (such as intubating in awkward positions, dangerous situations or in a moving vehicle), ability of the provider to sense the need for intubation (e.g., medics have been previously been shown to be able to reliably predict whether someone has aspirated [52]), and unmeasured injury status.

The instrumental variable used was the rate of PHI per metropolitan statistical area (MSA). We computed this variable from the same dataset as our analyses data. However, for each patient, the level of PHI use in that patient's MSA was calculated after excluding that patient from both the numerator and the denominator. The IV was expected to be predictive of PHI intubation use for individual patients– higher use of this procedure in an MSA would be associated with higher likelihood of the index patient undergoing PHI if treated in that region. The rates are also expected to be independent of the potential outcomes in the overall target population, as they are mostly driven by system-level resources and practice guidelines. However, among patients receiving PHI, patients in high PHI use area may have different levels of unobserved confounders than those living in a low use area. This creates dependence between the instrumental variable and the unobserved confounders among those receiving PHI, even if they are independent in the overall population. It treatment effects are heterogeneous over these unobserved confounders, the situation of essential heterogeneity arise, and the traditional IV effect estimates the effect for a marginal group of patients with a very specific level of the unobserved confounder.

Statistical Analysis

First, a naïve logistic regression analysis, controlling for observed levels of confounders, was run to study the adjusted effects of PHI. Various goodness of fit tests were used to ensure that the specification fit the data well.

Next, a traditional two-stage residual inclusion (2SRI) approach was applied using instrumental variables [53]. In the first stage, a logistic regression was used to predict the propensity to undergo PHI as a function of both observed confounders and the instrumental variables. A residual was computed by subtracting that predicted propensity score ($\hat{P}(Z,X)$) from the treatment indicator. In the second stage, another logistic regression was used to model the death indicator as a function of the PHI, observed confounders and the residual computed in the first stage. The treatment effect was computed based on the difference in predicted probability of death between PHI and no-PHI.

Finally, the local instrumental variables (LIV) approach was employed. In the LIV approach, the logistic outcome regression was run on all covariates (X), the estimated propensity score $\hat{P}(Z,X)$, the interaction of propensity score with all covariates, and a polynomial on the propensity score, $K(\hat{P};d)$:

$$E(Y) = Logit^{-1}(\beta_0 + X \cdot \beta_1 + X \cdot \hat{P} \cdot \beta_2 + K(\hat{P};d)) \quad (4)$$

The degree of polynomial, d, was selected based on likelihood-ratio tests between nested models with different degrees of polynomials for $\hat{P}(Z,X)$. The derivative $d\hat{E}(Y)/d\hat{P}(Z,X)$) of the final polynomial formulation was used as our LIV estimand to predict MTE(x,u_D):

$$MTE(x,u_D) = \left.\frac{d\hat{E}(Y)}{d\hat{P}}\right|_{\hat{P}=p}$$

The predicted values of the one minus the propensity score allow us to define the values of U_D over which MTE can be identified [54]. The larger the support of the propensity score, the bigger the set over which MTE can be recovered.

In order to study heterogeneity of effects over both the observed and unobserved characteristics, the dimensionality of X was reduced by using deciles of the estimated propensity of treatment choice based on observed confounders only.[4] These deciles are

[4] We ran a separate logistic regression for the treatment indicator on observed confounders X only to produce estimates for Pr(D =1|X=x). This is distinct from Pr(D =1|X=x, Z=z) used in the IV analyses.

denoted as η_q hereon, where q = 1,2,…,10. Thus, using our coefficient estimates from the above regression (equation (4)) MTE(η_q ,u_D) were estimated. The empirical conditional ATE(η_q) was estimated by averaging over ($u_D|\eta_q$) as ($u_D|\eta_q$)~ Uniform(Min{$\hat{P}(Z,X)|\eta_q$}, Max{$\hat{P}(Z,X)|\eta_q$}) by construction, while the empirical unconditional ATE was estimated using the empirical density of (η_q). Weights associated with TT and IV effect were computed and used to construct the respective treatment effect estimates. All analyses were weighted using probability weights and accounted for clustering of patients by hospitals. Standard errors for all the mean treatment effect parameters are estimated via 1000 clustered bootstrap replicates.

Results

There were 2169 patients in the NSCOT data who met all the inclusion criteria. Of them, 514 (23.7%) were intubated prior to arriving in the emergency department of the hospital where they received definitive care (PHI). Table 1 illustrates the differences in observed levels of potential confounders. Compared to non-PHI patients, patients who underwent PHI were on average 8 years younger, were less likely to be 65 years and older, were more commonly identified as racially White, and on Medicaid, but not Medicare. A higher proportion of patients who underwent PHI had prehospital shock, severe brain injury, but no Charlson comorbidity. PHI was also associated with the highest category of injury severity scores. There was no difference in the rate of PHI vs. no-PHI by gender or by admission to trauma center.

Based on the estimated propensity score for PHI receipts conditional on observed confounders and the instrumental variables, overlapping support across both treatment groups were found for 2069 patients (Figure 2). All of the following analyses were conducted on this sample.

Table 2 reports the estimated mean treatment effects based on different estimators. The naïve logistic regression analysis, controlling for observed levels of confounders produced a treatment effect estimate of 0.188 (se = 0.04, p < 0.001). It indicates that after controlling for observed confounders, on average, PHI increases in-patient mortality by 18.8% compared to no-PHI.

Next, the instrumental variable methods were evaluated. The instrumental variable was found to be a significant predictor (p < 0.0001) of PHI. In order to explore whether the observed confounders distributed uniformly over the levels of the instrumental variables (a necessary test), we predicted the propensity of PHI as a function of IV only and then compared the levels of observed confounders above and below the median of this predicted propensity. Because a

good instrument does not affect outcomes directly, it should considerably reduce the imbalance in levels of all confounders across the propensity median compared to the imbalances across treatment receipts. Table 1 reports the p-values for these comparisons on observed confounders and shows that the imbalance across levels of observed confounders were drastically reduced. Though not sufficient, this necessary test supports our theoretical assumptions that the imbalances in the levels of unobserved confounders should also be reduced across predicted propensity scores and the instrument should not directly affect outcomes.

The 2SRI approach estimated the treatment effects to be 0.029 (se = 0.08, p = 0.72) (Table 2). This estimate was lower than what was obtained via the naïve logistic regression. It says that the casual effect of PHI compared to no-PHI will increase in-patient mortality by 3%, which is not statistically significant from zero. This IV effect, however, may not correspond to any interpretable mean treatment effect parameter if essential heterogeneity is present. To explore this, the local instrumental variable approach was used.

Figure 3 shows how the marginal treatment effects, estimated using the LIV approach, vary jointly over the levels of observed confounders (η_q) and the latent dimension of unobserved confounding (U_D). Note that the joint support for (η_q,U_D) is not identified for low η_q & U_D and also for high η_q & U_D. This is not entirely a limitation of the data or the instrumental variable and rather has a strong behavioral flavor. Assuming that medics make the decision to perform PHI in the trauma patient, then if both the observed and unobserved levels of confounders make individual patients less likely to receive PHI, the medics are less likely to respond to an exogenous stimulus (instrument) that push them towards using PHI on these patients. In a similar vein, if characteristics make individuals highly likely to receive PHI, the medics are less likely to respond to an exogenous stimulus to withhold PHI for these patients. In other words, in these settings, there are no *marginal* patients who are at the precipice of choice based on their observed and unobserved characteristics and a perturbation in an instrument can determine their treatment receipt. To most extent, unless new information or changes in system level incentives modifies the delivery of these treatments, available data cannot be used to estimate a casual treatment effect in these patients.[5] We focus our attention to the patients where such comparisons can be made.

[5] This concept of the absence of the marginal patients is somewhat similar, though not identical to the concepts of *never takers* and *always takers* in Angrist et al [15].

First, treatment effects heterogeneity was explored based on observed confounders. Figure 3 reports the average treatment affects over levels of observed confounding. It shows that as the propensity to receive PHI based on observed characteristics increased, treatment effects became smaller and ultimately negative (e.g. ATE(η_{10}) = -0.07 (0.02), p<0.001). This indicates positive self-selection based on these characteristics. In fact, the proportion receiving PHI was 40% among those who belonged to η_{10} compared to 14.5% in the other deciles of observed confounders (p< 0.001). Focusing only on subgroup analyses based on observed characteristics, it could be concluded that 89% of the patients belonged to deciles η_1 to η_9 and, on average, would experience significant increase in mortality had they undergone PHI versus no-PHI. These patients are said to constitute the *extensive margin* [26] of the comparison. However, focusing simultaneously on the heterogeneity dimension based on unobserved characteristics, it is found that the extensive margin may comprise of only 59% of the patients if only the point estimate at each margin is considered. That is, even if the conditional average effect is positive for patients in η_1 to η_9, not all patients within that group would have a positive treatment effect.

Next, effect heterogeneity over unobserved confounders was explored. Conditional on a level of observed confounders, the profile of MTEs over U_D is analogous to the stylized situation depicted in Figures 1(c) and (d). We found evidence of essential heterogeneity. Much like the observed confounders, unobserved confounders that made patients highly likely to undergo PHI are also associated with negative treatment effects or decreased mortality from PHI versus no-PHI. Similarly, levels of unobserved confounders that made patients less likely to undergo PHI are associated with positive treatment effects or increased mortality from PHI versus no-PHI. This provides evidence on *positive* self-selection behavior based on unobserved confounders; therefore, in practice, the effect of PHI on patients who were undergoing PHI is determined by the combined positive self-selection based on different confounders. Consequently, the mean treatment effect parameters estimates were found to be substantially different that either the naïve or the traditional IV estimates (Table 2). The average treatment effect (ATE) was estimated to be 0.074 (se = 0.01, p <0.001), while the effect on the treated (TT) was estimated to be -0.079 (se = 0.09, p = 0.38). It indicates that had PHI been conducted on all trauma subjects, the average inpatient mortality rate would have increased by 7.4% points when compared to the situation where all trauma patients undergo no PHI. However, this estimate may be misleading because such an all-or-none approach is hardly a pragmatic comparison. Perhaps, more relevant is the estimate for the effect on the treated (TT), which indicates that those who are undergoing PHI in practice would have increased their average mortality rate by

7.9% had they not been subjected to PHI, although this estimate does not reach statistical significance.

In a more nuanced exploratory analysis, observed factors associated with the observed characteristics-based *extensive margin* were explored using univariate comparisons.[6] Basically, the levels of individual *X*s were compared between patients who belonged to η_{10} versus those who belonged to η_1 to η_9. Table 3 presents these results. The most significant factors associated with harms versus benefits were found to be older age, non-Hispanic Black race/ethnicity, and Medicare insurance. As expected and evident from Table 1, many of these characteristics also predicted non-receipt of PHI proving strong evidence of positive self-selection in practice based on these characteristics. Interestingly, patients with severe head injury and those with ISS scores>34 had the clearest benefit from PHI. Table 3 also reports characteristics associated with increased mortality with PHI among those who were undergoing PHI. Compared to all patients, among patients receiving PHI, pre-hospital shock, admission to trauma center and comorbidities were no longer significant predictors of harm, indicating that selection on these covariate dimensions were already taking place. On the other hand, age, severe head injury status and those with ISS scores>34 were some of the dimensions along which selection could be further enhanced.

Finally, whether areas with higher PHI use were associated with less positive self-selection (i.e., more overuse) was studied. That is, we wanted to know whether subjects who received PHI in lower use regions were less likely to be harmed by PHI than those who received PHI is higher use regions. Regions where PHI use rates were lower than the median rate were compared to regions where they were higher. Table 4 reports the results, as well as the estimated risk of harm that is the probability of belonging to deciles η_1 to η_9. Compared to the risk of harm by PHI among the entire population, the risk of harm by PHI among PHI recipients is significantly reduced in low use areas ($p < 0.001$), but not in high use areas ($p = 0.12$). These results provide strong evidence of over-use of PHI in high-use regions.

[6] Ideally, one can develop prediction algorithms using multivariate analysis on such information. However, our sample size was too small to develop and validate such algorithm. We, therefore, present some exploratory analyses leaving room for future work in this area.

Conclusions

With the increased investment in developing large observational studies for comparative effectiveness studies, methods that produce valid and interpretable results are in great demand. Moreover, as the field of CER moves towards a patient-centered paradigm, understanding heterogeneity in comparative effects becomes crucial. In this paper, we have highlighted the recent development of instrumental variable methods to address such challenges. We applied these methods towards a substantive problem of evaluating the comparative effectiveness of prehospital intubation versus no prehospital intubation on inpatient mortality in trauma patients.

One critical step, in trying to align these analyses to decision making in practice, is to be able to measure variables/ confounders that are also readily available at the point of decision making, so that conditioning on them is pragmatic. We attempted to select variables that would be readily available for decision making from the start of care in the prehospital setting. Demographics are fairly easy to assess by observation. Calculation of the ISS score is time consuming and requires information not available to formally assess the extent of injury during the initial evaluation of the patient, but broad categorization of injury severity (mild, moderate, severe, near fatal) would be possible in the moments prior to definitive prehospital treatment. Furthermore, severe head injury with decreased level of consciousness is estimable, and prehospital shock (as defined by systolic blood pressure<90 mmHg) is measureable. It is less likely that sufficient history would be known to estimate the patient's comorbidities or pre-injury insurance status. The admission to a trauma center is likely a result of the decision making made at the scene, but may have been directed by preliminary reports of the level of injury and facilitated specialized prehospital care from the outset. As prehospital electronic data management improves, better capture of scene data may inform development of more specific mortality risk assessment. With real-time decision support, the prehospital provider may be able to improve their patient selection for PHI.

It is remarkable that for the most part, PHI is appropriately used, especially in low-use areas. Based on the prevailing concern for the safety of PHI due to difficulty with maintenance of intubation skills in low-PHI use prehospital systems, it is somewhat surprising to find that areas of limited PHI use actually confer less risk to trauma patients due to PHI. One potential explanation may be that in low-use regions, the number of prehospital providers trained to perform intubation is also low and hence these medics may accumulate intubating experience at a higher rate than those in high-use regions, making them more efficient with the intubation procedure [55].

Additionally, the difference in the level of PHI risk among PHI recipients between low and high-PHI regions may be explained by patient selection for PHI. Our analysis also reveals that those patients without head injury, with lower injury severity and without prehospital shock had a higher likelihood of significant harm from PHI. Not surprisingly, areas with high rates of PHI use had a lower risk reduction due to positive self-selection, presumably because the indications for PHI were liberalized. Ultimately, the positive effects of PHI are more pronounced when it is reserved for only the most critically ill patients that survive to the emergency department – in particular those with severe head injury; when the indication for PHI is expanded beyond this, at best, PHI confers no survival advantage, and for the relatively uninjured, the risk of PHI may outweigh the risk of mortality without PHI.

One limitation of our study is that the prehospital death rate is unknown; patients who were declared dead at the scene, en route, or within 30 minutes of arrival if they arrived at the hospital without signs of life were excluded from the NSCOT study. If PHI has a differential effect on pre-hospital mortality than no PHI, then focusing analysis on patients who reach the hospital may provide biased assessment of PHI. It should also make the case mix of patients who reach the hospital alive in low-use regions different from that in high-use regions, thereby invalidating the IV. However, we did not find any evidence of such differential case-mix in our observed data. Moreover, we believe that even if there may be a differential effect of PHI on pre-hospital mortality, it is likely to be quite small as the overall rate of pre-hospital mortality in patients with severe trauma is small (~3% [56]).

Overall, we believe that the methods highlighted in this paper can provide a rich set of tools for researchers to explore hypothesis on heterogeneity in treatment effects. Obviously, it is necessary to replicate these results before they are implemented in practice. However, such results can provide informative priors for designing confirmatory trials/studies in this setting, thereby making the link between information generation and decision making more efficient, and in line with accomplishing the goals of CER.

Appendix:

$$E(Y \mid Z = z, X = x) = E(DY_1 + (1-D)Y_0 \mid Z = z, X = x)$$

$$= E(Y_0 \mid X = x) + E(D(Y_1 - Y_0) \mid Z = z, X = x)$$

$$= E(Y_0 \mid X = x) + \Pr(D = 1 \mid Z = z, X = x) \cdot E((Y_1 - Y_0) \mid D = 1, X = x)$$

$$= E(Y_0 \mid X = x) + \Pr(D = 1 \mid Z = z, X = x) \cdot \frac{\int_{S(z,x)=1-P(z,x)}^{1} E((Y_1 - Y_0) \mid U_D = u, X = x) du}{\int_{S(z,x)=1-P(z,x)}^{1} du}$$

$$= E(Y_0 \mid X = x) + \int_{S(z,x)=1-P(z,x)}^{1} E((Y_1 - Y_0) \mid U_D = u, X = x) du$$, where the last equality follows as $D =$ ($U_D > S(z,x)$) and therefore, $\Pr(D = 1 \mid Z = z, X = x) = \int_{S(z,x)=1-P(z,x)}^{1} du$.

Let $E((Y_1 - Y_0) \mid U_D = u, X = x) = g(u, x)$, where $g(U, X)$ is some non-parametric function of X and U. Now, if we take the rate of change of the mean outcome Y with respect to the probability of receiving treatment, P:

$$\frac{\partial}{\partial P(z,x)} E(Y \mid Z = z, X = x)\Big|_{\hat{P}(x,z)=p(x,z)} = \frac{\partial}{\partial P(z,x)}\left[\int_{S(z,x)}^{1} g(u,x) du\right] = \frac{\partial S(z,x)}{\partial P(z,x)} \cdot \frac{\partial}{\partial S(z,x)}\left[\int_{S(z,x)}^{1} g(u,x) du\right]$$

$$= \frac{\partial}{\partial S(z,x)}\left[\int_{0}^{S(z,x)} g(u,x) du\right] = g(S(z,x), x) = \text{MTE}(x, u_D)$$

Therefore, the LIV estimand comprises of the partial derivative of the control function for Y with respect to P, and evaluating this derivative by setting $P = p$, where p is a specific estimated value of $\hat{P}(Z, X)$ in the data, to obtain an estimate of MTE(U_D=1-p,x).

The formal proof of consistency for this estimator can be found in Heckman et al. (2006).

Figure 1: A stylized illustration of non-essential (a & b) and essential heterogeneity (c) and estimation of marginal treatment effect (MTEs) using the local instrumental variable (LIV) approach (d).

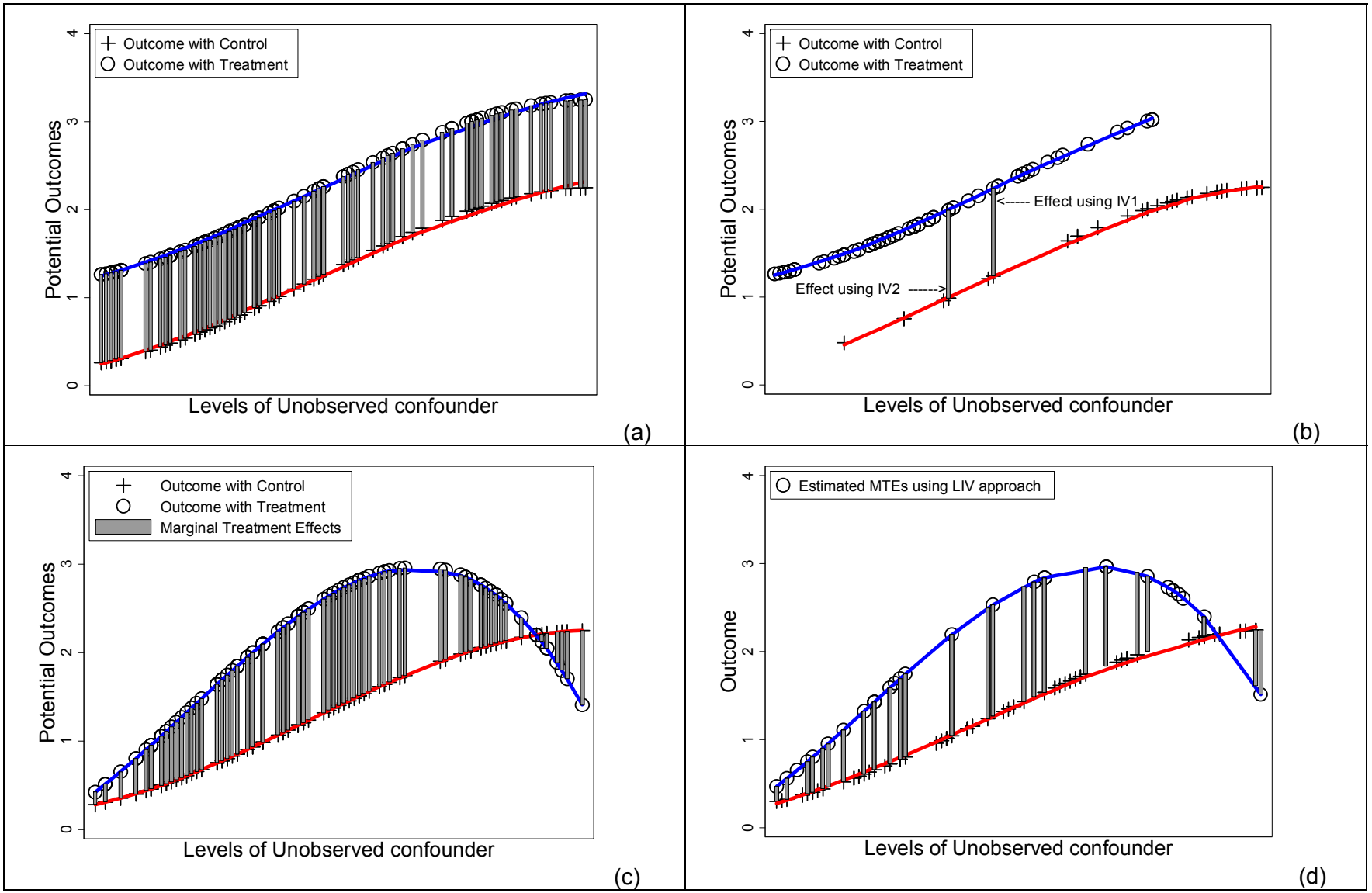

Figure 2: Distribution of estimated propensity score to receive prehospital intubation for receivers and non-receivers of prehospital intubation, conditional on observed confounders and instrumental variables.

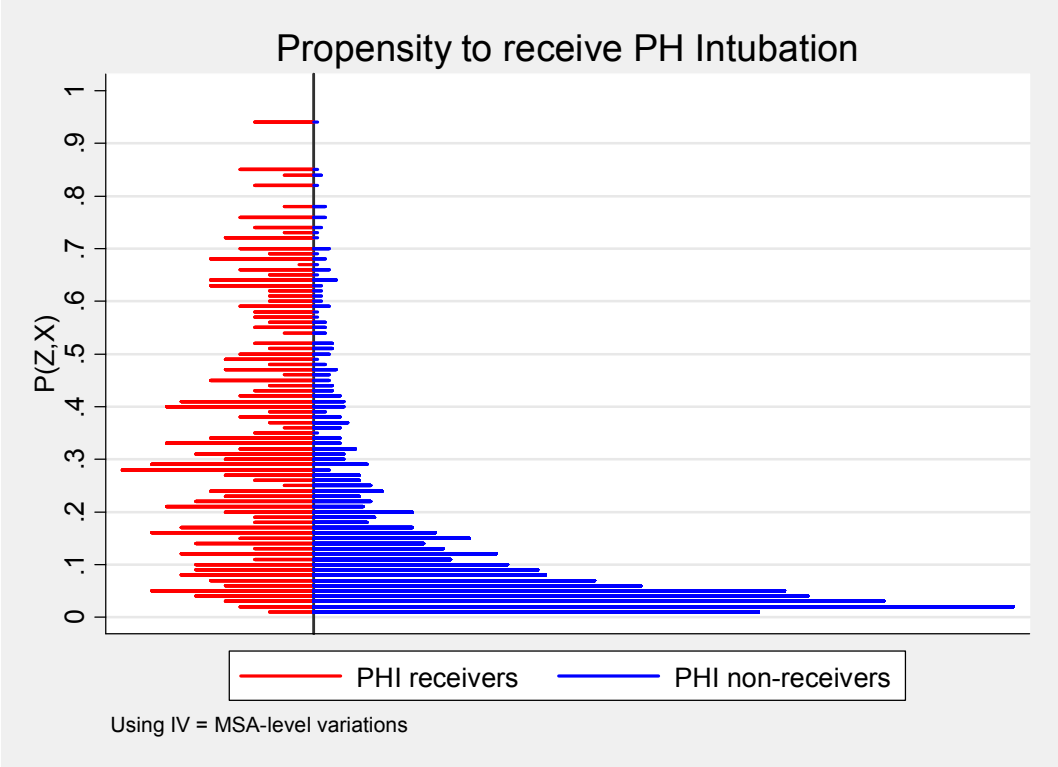

Figure 3: (a): Marginal treatment effect profile jointly across the deciles of propensity to receive prehospital intubation based on ***observed*** confounders (η_q; q = 1,..,10) and the propensity to receive prehospital intubation based on ***unobserved*** confounders (U_D). (Black dots indicate the specific margins where treatment effects are significantly different that zero at 5% level; * significance at 5% level)

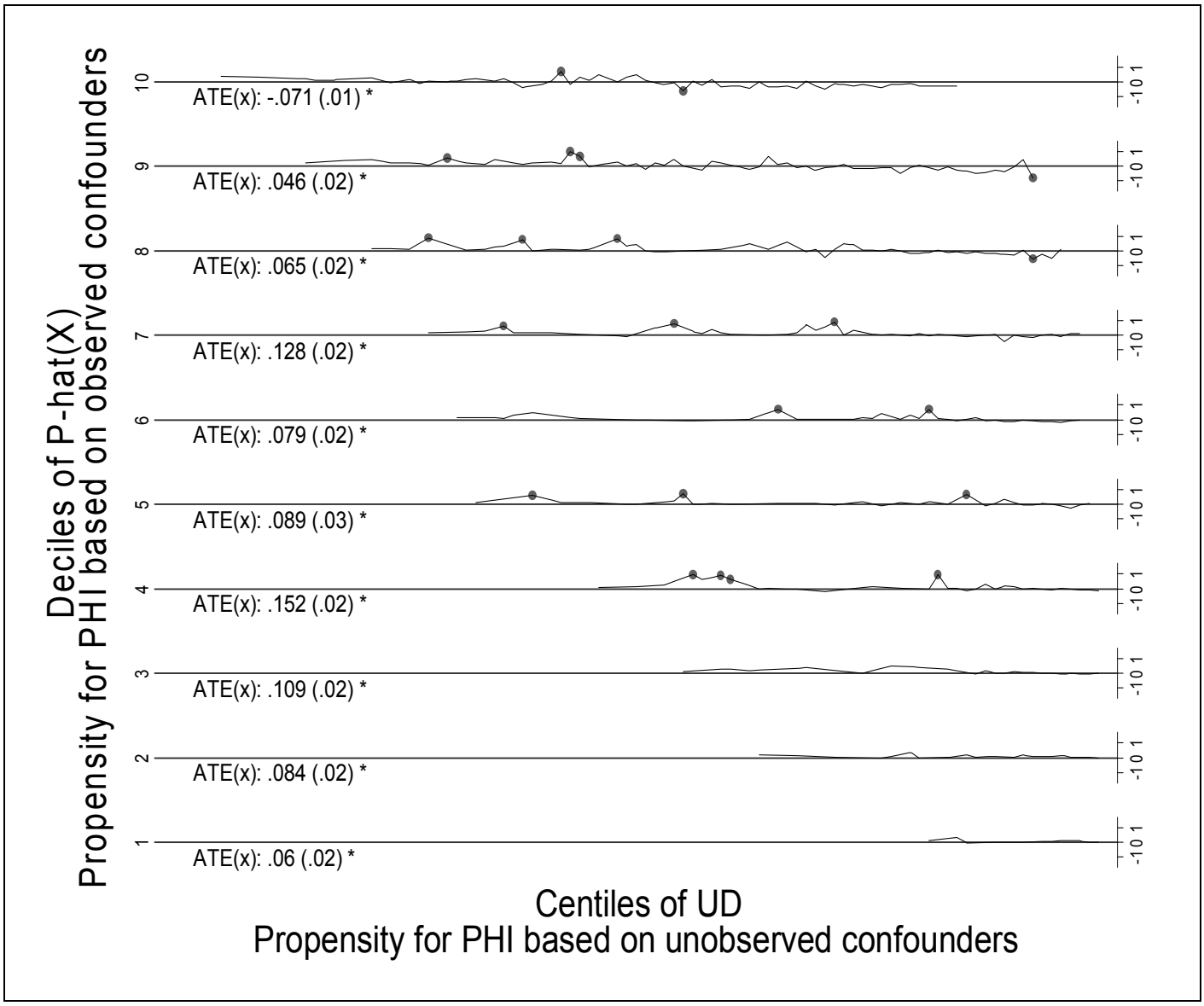

Table 1: Characteristics of observed confounders across treatment arms

Characteristics	No Prehospital Intubation N = 1655	Prehospital Intubation N = 514	p-value	p-value across IV levels*
	% or Mean (sd)	% or Mean (sd)		
Age (in years)	47 (20)	39 (18)	< 0.001	0.44
Age 18 – 24	22.9%	33.0%		
Age 25 - 34	17.6%	24.1%		
Age 35 – 44	18.9%	14.1%		
Age 45 – 54	15.2%	13.1%		
Age 55 – 64	9.9%	9.1%		
Age > 64	26.1%	12.3%	< 0.001	0.26
Female	26.4%	29.1%	0.27	0.25
Race/Ethnicity:				
Hispanic	20.8%	14.7%		
Non Hispanic White	49.3%	67.8%		
Non Hispanic Black	29.9%	17.5%	0.07	0.09
Pre-injury Insurance:				
Uninsured	32.5%	29.9%		
Medicare	17.7%	8.9%		
Private only	38.3%	41.5%		
Medicaid and Other	11.5%	19.7%	0.02	0.05
Prehospital shock	13.1%	27.2%	<0.001	0.70
Severe head injury	49.3%	76.3%	<0.001	0.10
Admitted to a Trauma Center	86%	88%	0.12	0.64
Charlson comorbidity				
No comorbidity	74.0%	86.1%		
1 comorbidity	14.6%	5.7%		
2 comorbidities	5.2%	4.5%		
3+ comorbidities	6.3%	3.7%	0.002	0.05
ISS Score quartiles:				
16-24	57.1%	28.6%		
25-34	33.9%	39.4%		
>34	9.0%	32.0%	< 0.001	0.81
In-hospital death	9%	31%	< 0.001	< 0.001

* Comparison of covariate levels between above and below median values of the IV. ISS=Injury Severity Score.

Table 2: Treatment effect estimates of prehospital intubation compared to no prehospital intubation on inpatient mortality, across alternative methods.

Estimator	Mean effect (se) [p-value]
Unadjusted	**0.220 (0.02) [<0.001]**
Regression (Logistic)	**0.188 (0.03) [<0.001]**
IV (2stage residual inclusion)	0.029 (0.08) [0.72]
LIV-based estimates	
IV-effect	0.011 (0.08) [0.89]
Average Treatment Effect (ATE)	**0.074 (0.01) [<0.001]**
Effect on the Treated (TT)	- 0.079 (0.09) [0.38]

Table 3: Differences in levels of observed confounders across groups with significant harm (Pr(η_1 to η_9)) versus significant benefit (Pr(η_1 to η_9)) on average.

	ALL PATIENTS			PHI RECIPIENTS		
Characteristics	Significant benefit from PHI N = 200	Significant harm from PHI N = 1869	p-value	Significant benefit from PHI N = 123	Significant harm from PHI N = 302	p-value
	% or Mean (sd)			% or Mean (sd)		
Age (in years)	29 (18)	41 (20)	<0.001	29 (18)	37 (20)	<0.001
Age 18 – 24	47.9%	24.1%		47.1%	34.1%	
Age 25 – 34	29.8%	18.9%		29.6%	18.7%	
Age 35 – 44	12.9%	20.4%		15.8%	15.6%	
Age 45 – 54	6.3%	11.5%		4.1%	12.5%	
Age 55 – 64	2.0%	9.4%		1.7%	9.8%	
Age > 64	1.0%	16.0%	<0.001	1.7%	9.3%	0.035
Female	33.1%	28.7%	0.174	34.0%	29.5%	0.693
Race/Ethnicity:						
Hispanic	21.6%	23.3%		20.5%	17.1%	
White	75.6%	51.0%		77.3%	58.9%	
Black	2.8%	25.7%	< 0.001	2.2%	24.0%	0.020
Pre-injury Insurance:						
Uninsured	31.3%	32.2%		32.1%	29.9%	
Medicare	2.9%	17.2%		3.5%	11.0%	
Private only	45.3%	41.5%		33.3%	44.8%	
Medicaid and Other	20.5%	9.1%	<0.001	31.1%	14.3%	0.025
Prehospital shock	34.8%	14.6%	0.004	28.9%	20.4%	0.257
Severe head injury	90.6%	52.2%	<0.001	94.2%	68.9%	<0.001
Admitted to a Trauma Center	92.9%	87.1%	0.003	89.5%	85.7%	0.537
Charlson comorbidity						
No comorbidity	88.1%	79.8%		88.8%	85.1%	
1 comorbidity	7.5%	7.3%		6.7%	5.7%	
2 comorbidities	3.1%	5.3%		3.2%	4.9%	
3+ comorbidities	1.3%	7.6%	0.017	1.3%	4.3%	0.473
ISS Score quartiles:						
16-24	2.5%	59.2%		0.4%	46.0%	
25-34	35.6%	33.2%		34.8%	39.6%	
>34	61.9%	7.6%	<0.001	64.8%	14.4%	<0.001

Table 4: Comparing low PHI use regions to high PHI use regions for degree of positive self-selection of PHI use.

Proportion who belong to subgroups where PHI is estimated to produce significant harm on average [p-value]	Low PHI use regions (< Median rate)	High PHI-use regions (≥ Median rate)	p-value for Difference
Among all patients*	85.1% [<0.001]	87.9% [<0.001]	0.34
Among patient receiving PHI**	62.5% [<0.001]	80.0% [<0.001]	0.08
Risk reduction due to positive self-selection	-26% [<0.001]	-9% [0.12]	0.13

*Pr(η_1 to η_9); **Pr(η_1 to η_9|D=1);

References

[1] Tunis SR, Benner J, McClellan M. Comparative effectiveness research: Policy context, methods development, and research infrastructure. *Statistics in Medicine* 2010; **29**: 1963-1971.

[2] Wang HE, Peitzman AB, Cassidy LD, Adelson PD, Yealy DM. Out-of-hospital endotracheal intubation and outcome after traumatic brain injury. *Annals of Emergency Medicine* 2004; **44(5)**: 439-50.

[3] Evans HL, Zonies DH, Warner KJ, et al. Timing of intubation and ventilator-associated pneumonia following injury. *Archives of Surgery* 2010;**145(11)**:1041-1046.

[4] 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.

[5] Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Annals of Internal Medicine* 2002; **137**:273-84.

[6] 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**:766-79.

[7] Bernard SA, Nguyen V, Cameron P, Masci K, Fitzgerald M, Cooper DJ, et al. Prehospital rapid sequence intubation improves functional outcome for patients with severe traumatic brain injury: a randomized controlled trial. *Annals of Surgery* 2010**; 252(6)**, pp. 959-65

[8] Stock JH, Trebbi F. Who invented instrumental variable regression? *Journal of Economic Perspective* 2003; **17(3)**: 177-194.

[9] McClellan M, McNeil B, Newhouse J. Does More Intensive Treatment of Acute Myocardial Infarction Reduce Mortality? *JAMA* 1994; **272(11)**:859-866.

[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. *Journal of Clinical Oncology* 2001; **19(4)**: 1064-1070.

[11] 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 Services Research* 2003; **38(6)**, Part I: 1385-1402.

[12] 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 Economics* 2003; **12**: 171-186.

[13] 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; 297:278-85.

[14] Amemiya T. The non-linear two-stage least squares estimator. *Journal of Econometrics* 1974; 105-110.

[15] Angrist J, Imbens G, Rubin D. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 1996; **91**: 444-455.

[16] Heckman J. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* 1997; **32(3)**: 441-462.

[17] Greenfield S, Kravitz R, Duan N, Kaplan SH. Heterogeneity of treatment effects: implications for guidelines, payment, and quality assessment. *American Journal of Medicine* 2007; **120**(4):S3--S9.

[18] Imbens G, Angrist J. Identification and estimation of local average treatment effects. *Econometrica* 1994; **62(2)**: 467-475.

[19] Heckman JJ. Comments on Angrist, Imbens, and Rubin: Identification of Causal Effects Using Instrumental Variables. Journal of American Statistical Association 1996; **91**: 434.

[20] Heckman JJ, Vytlacil EJ. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences* 1999; **96(8)**: 4730-34.

[21] Heckman J. and Vytlacil E. Structural equations, treatment effects and econometric policy evaluation. *Econometrica* 2005; **73(3)**: 669-738.

[22] Heckman JJ, Urzua S, Vytlacil E. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* 2006; **88(3)**: 389-432.

[23] Vanness DJ, Mullahy J. *Perspectives on Mean-based Evaluation of Health Care*. In Jones A. (Eds) The Elgar Companion to Health Economics, Edward Elgar Publishing: Cheltenham; 2006.

[24] Basu A. Individualization at the heart of comparative effectiveness research: The time for i-CER has come. *Medical Decision Making* 2009; **29(6)**: N9-N11.

[25] Basu A, Jena A. Philipson T. Impact of comparative effectiveness research on health and healthcare spending. 2010 NBER Working Paper No. w15633. *Journal of Health Economics 2011;* In press.

[26] Basu A. Economics of individualization in comparative effectiveness research and a basis for a patient-centered healthcare. 2011 *NBER Working Paper* No. w16900. *Journal of Health Economics* 2011; **30(3)**; 549-559.

[27] 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 Economics* 2007; 16(11): 1133 -1157.

[28] Basu A. Estimating decision-relevant comparative effects using instrumental variables. *Statistics in Biosciences* 2011. In press.

[29] Bratton SL, Chestnut RM, Ghajar J, McConnell Hammond FF, Harris OA, Hartl R, et al. Guidelines for the management of severe traumatic brain injury. I. Blood pressure and oxygenation, *J Neurotrauma* 2007; **24 Suppl 1**: S7-13

[30] Davis DP, Fakhry SM, Wang HE, Bulger EM, Domeier RM, Trask AL, et al. Paramedic rapid sequence intubation for severe traumatic brain injury: perspectives from an expert panel, *Prehosp Emerg Care* 2007; **11(1)**: 1-8

[31] Davis DP, Hoyt DB, Ochs M, Fortlage D, Holbrook T, Marshall LK, et al. The effect of paramedic rapid sequence intubation on outcome in patients with severe traumatic brain injury, *Journal of Trauma* 2003; **54(3)**: 444-53

[32] Fakhry SM, Scanlon JM, Robinson L, Askari R, Watenpaugh RL, Fata P, et al. Prehospital rapid sequence intubation for head trauma: conditions for a successful program. *Journal of Trauma* 2006; **60(5)**: 997-1001.

[33] Ang DN, Rivara FP, Nathens A, Jurkovich GJ, Maier RV, Wang J, et al. Complication rates among trauma centers, *Journal of the American College of Surgery* 2009; **209(5)**: 595-602

[34] Cooper Z, Rivara FP, Wang J, MacKenzie EJ, Jurkovich GJ. Withdrawal of life-sustaining therapy in injured patients: variations between trauma centers and nontrauma centers, *Journal of Trauma* 2009; **66(5)**: 1327-35

[35] Davydow DS, Zatzick DF, Rivara FP, Jurkovich GJ, Wang J, Roy-Byrne PP, et al. Predictors of posttraumatic stress disorder and return to usual major activity in traumatically injured intensive care unit survivors, *Gen Hosp Psychiatry* 2009; **31(5)**: 428-35

[36] MacKenzie EJ, Weir S, Rivara FP, Jurkovich GJ, Nathens AB, Wang W, et al. The value of trauma center care, *Journal of Trauma* 2010; **69(1)**: 1-10

[37] Nathens AB, Rivara FP, Wang J, Mackenzie EJ, Jurkovich GJ. Variation in the rates of do not resuscitate orders after major trauma and the impact of intensive care unit environment, *Journal of Trauma* 2008; **64(1)**: 81-8; discussion 8-91

[38] Sorensen MD, Wessells H, Rivara FP, Zonies DH, Jurkovich GJ, Wang J, et al. Prevalence and predictors of sexual dysfunction 12 months after major trauma: a national study, *Journal of Trauma* 2008; **65(5)**: 1045-52; discussion 52-3

[39] Zatzick D, Jurkovich GJ, Rivara FP, Wang J, Fan MY, Joesch J, et al. A national US study of posttraumatic stress disorder, depression, and work and functional outcomes after hospitalization for traumatic injury. *Annals of Surgery* 2008; **248(3)**: 429-37

[40] Chaiwat O, Lang JD, Vavilala MS, Wang J, MacKenzie EJ, Jurkovich GJ, et al. Early packed red blood cell transfusion and acute respiratory distress syndrome after trauma, *Anesthesiology*. **2009; 110(2)**, pp. 351-60

[41] Efron DT, Sorock G, Haut ER, Chang D, Schneider E, Mackenzie E, et al. Preinjury statin use is associated with improved in-hospital survival in elderly trauma patients, *Journal of Trauma* 2008; **64(1)**: 66-73.

[42] Bulger EM, Nathens AB, Rivara FP, MacKenzie E, Sabath DR, Jurkovich GJ. National variability in out-of-hospital treatment after traumatic injury, *Annals of Emergency Medicine* 2007; **49(3)**, pp. 293-301

[43] McFadden D. *Conditional logit analysis of qualitative choice behavior.* In: Zarembka P (ed) Frontiers in Econometrics. Academic Press, New York; 1973.

[44] McFadden D. *Econometric models of probabilistic choice.* In: Manski CF and McFadden D (eds) Structural Analysis of Discrete Data with Econometric applications, Cambridge. LIA: MIT Press; 1981.

[45] Björklund A, Moffitt R. The estimation of wage gains and welfare gains in self-selection. *Review of Economics and Statistics* 1987; **69(1)**: 42-49.

[46] Heckman JJ, Vytlacil E. *Econometric evaluation of social programs (part II)*. In J. Heckman and E. Leamer (Eds.) Handbook of Econometrics vol 6, Elsevier: Amsterdam; 2007

[47] MacKenzie EJ, Rivara FP, Jurkovich GJ, Nathens AB, Frey KP, Egleston BL, et al. A national evaluation of the effect of trauma-center care on mortality, *N Engl J Med* 2006; **354(4)**: 366-78.

[48] MacKenzie EJ, Rivara FP, Jurkovich GJ, Nathens AB, Frey KP, Egleston BL, et al. The National Study on Costs and Outcomes of Trauma, *The Journal of Trauma* 2007; **63(6 Suppl)**: S54-67; discussion S81-6

[49] Dunham CM, Barraco RD, Clark DE, Daley BJ, Davis FE, Gibbs MA, et al. Guidelines for emergency tracheal intubation immediately after traumatic injury. *The Journal of Trauma: Injury, Infection, and Critical Care* 2003; **55(1)**:162–179.

[50] Warner KJ, Carlbom D, Cooke CR, Bulger EM, Copass MK, Sharar SR. Paramedic training for proficient prehospital endotracheal intubation. *Prehosp Emerg Care* 2010; **14(1)**:103-8.

[51] Wang HE, Seitz SR, Hostler D, Yealy DM. Defining the learning curve for paramedic student endotracheal intubation. Prehosp Emerg Care. 2005 Apr-Jun; 9(2):156-62.

[52] Vadeboncoeur TF, Davis DP, Ochs M, Poste JC, Hoyt DB, Vilke GM. The ability of paramedics to predict aspiration in patients undergoing prehospital rapid sequence intubation. *Journal of Emergency Medicine* 2006; **30(2)**: 131-6.

[53] Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 2008; **27(3)**:531-543.

[54] Heckman JJ, Vytlacil E. *Local instrumental variables*. In C. Hsiao, K. Morimue, and J.L. Powell (Eds.) Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in the Honor of Takeshi Amemiya, Cambridge University Press: New York, 2001: 1-46.

[55] Garza AG, Gratton MC, Coontz D, Noble E, Ma OJ. Effect of paramedic experience on orotracheal intubation success rates. *The Journal of Emergency Medicine* 2003; **25(3)**: 251-256.

[56] Sollid SJM, Lossius HM, Soreide E. Pre-hospital intubation by anaesthesiologists in patients with severe trauma: an audit of a Norwegian helicopter emergency medical service. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 2010; **18:** 30.