NBER WORKING PAPER SERIES

VIEWPOINT: THE HUMAN CAPITAL APPROACH TO INFERENCE

W. Bentley MacLeod

Working Paper 22123
http://www.nber.org/papers/w22123

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2016

This paper is based upon the lecture given at the Canadian Economics Association Meetings, Toronto, May 2015. I am grateful to Jonathan Cohen, Janet Currie, Angus Deaton, Sebastein Seung and Jacques Thisse for helpful discussions. I am particularly grateful to Elliott Ash, Daniel Deibler and Xuan Li for invaluable research assistance on this project, and to Charles Beach, President of the Canadian Economics Association, for inviting me to talk about this topic. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Viewpoint: The Human Capital Approach to Inference
W. Bentley MacLeod
NBER Working Paper No. 22123
March 2016, Revised June 2016
JEL No. C1,I1,I18,J24

ABSTRACT

The purpose of this essay is to discuss two approaches to inference, and how "human capital" can provide a way to combine them. The first approach, ubiquitous in economics, is based upon the Rubin/Holland potential outcomes model and relies upon randomized treatment to measure the causal effect of choice. The second approach, widely used in the pattern recognition and machine learning literatures, assumes that choice conditional upon current information is optimal (or at least high quality), and then provides techniques to generalize observed choice to new cases. The "human capital" approach combines these methods by using observed decisions by experts to reduce the demensionality of the feature space and allow the categorization of decisions by their propensity score. The fact that the human capital of experts is heterogeneous implies that errors in decision making are inevitable. Moreover, under the appropriate conditions, these decisions are random conditional upon the propensity score. This in turn allows us to identify the conditional average treatment effect for a wider class of situations than would be possible with randomized control trials. This point is illustrated with data from medical decision making in the context of treating depression, heart disease, and adverse childbirth events.

W. Bentley MacLeod
Department of Economics
Columbia University
420 West 118th Street, MC 3308
New York, NY 10027
and NBER
bentley.macleod@columbia.edu

"False facts are highly injurious to the progress of science, for they often endure long; but false views, if supported by some evidence, do little harm, for everyone takes a salutary pleasure in proving their falseness; and when this is done, one path towards error is closed and the road to truth is often at the same time opened." - Charles Darwin, *Descent of Man,* Vol 2, Chapter 2, 1871.

1 Introduction

There are two distinct approaches in empirical labor economics. The first approach addresses the indentification problem that arises when individuals self-select into different observed treatments or choices by either explicitly randomizing treatments/choices in the context of an experiment (Charness and Kuhn (2011) and List and Rasul (2011)), or through the use of a natural experiment that allows for an instrumental variables strategy (Angrist et al. (1996) and Angrist and Krueger (1999)). The second approach uses structural models that assume individuals make utility maximizing decisions within a well defined environment, and then proceeds to measure the value of the unknown parameters. A classic example of this is the well known Roy (1951) model, where we know that the model can only be identified under strong assumptions (Heckman and Honore (1990)).

In this paper I review some recent work that combines these perspectives to provide a way to extend the scope of randomization to environments where randomized control trials are not possible, either due to the problem of constructing an adequate subject pool, or because the number of cases to be considered is simply too large. The fact that randomized trials are limited by their costs has long been recognized. Fisher (1936) was the early leader in the field, with early work tackling the problem of improving agricultural production in developed (Yates (1933); Bose and Mahalanobis (1938)) and developing countries (Bose and Mahalanobis (1938)). Such experiments can take many years, and thus it was understood early on that one could could not rely only upon experimental methods. For example, Mahalanobis (1944) provides a wonderful discussion of the survey techniques he developed to supplement experimental studies of Indian agriculture.

The rise of experimental economics may be attributed to the combination of many new game theoretic ideas developed in recent decades, combined with the fact they these ideas could be explored at a relatively low cost using college students as subjects. Moreover, there was an increased awareness of some of the stringent conditions needed to ensure the causal identification of an intervention (Holland (1986) and Imbens and Rubin (2011)). Thus we have also seen a large increase in the use of field experiments that measure the effect of treatment using using realistic interventions.[1] These experiments increase the external validity of the results relative to laboratory experiments. However, they are limited in both the size of the monetary rewards than can be used,

[1]See for example List and Rasul (2011) and Banerjee and Duflo (2009).

and the period of time over which it is feasible to run a field experiment. As a consequence, Deaton (2010) has observed that many questions of interest and importance cannot be studied with purely experimental techniques.

One of these areas is expert decision making, particular by physicians. Medical decision making is a particularly interesting case because RCTs (randomized control trials) are widely used to explore the efficacy of medical treatments.[2] The next section briefly reviews the Rubin/Holland potential outcomes framework and shows that it has performed rather poorly in determining the appropriate intervention for the treatment of depression.[3] This is a nice example for a number of reasons. First, the treatment of depression with medication is a multi-billion industry, funded in large part by health insurance. For example, in 2013 Abilify (Aripiprazole) was the *top* selling drug in the United States. It was initially approved for the treatment of schizophrenia, but is now used "off label" for a wide variety psychiatric conditions. In the absence of good clinical guidance, there may be a potentially large miss-allocation of resources (see Frank and McGuire (2000)). Second, subjects who face a high risk of suicide are, for ethical reasons, barred from participating in these studies for the treatment of depression, yet they are one of the prime beneficiaries of good treatment. Third, measuring the outcome of an intervention is difficult. In the case of depression one uses a survey instrument that may or may not be related to outcomes such as suicidality and labor market performance. Fourth, the response to the intervention is very heterogeneous. In the case of SSRIs (Selective serotonin reuptake inhibitors), the effect can vary from feeling slightly better to increased suicide risk. The challenge is to be able to predict for a given patient the likely consequence of treatment given his or her characteristics.

The heterogeneity in response presents a particular challenge. When a drug, such as say an antibiotic, is expected to be relatively safe, then the goal of an RCT is to measure it's effectiveness. With a large number of individuals one can obtain a good measure of the average treatment effect. The difficulty arises where there is heterogeneity in the sign of the treatment effect - it harms some individuals and not others. In that case if one ignores the heterogeneity, then the treatment effect from a trial might be zero, even though the drug is very effective (or dangerous) for some individuals. When the variability in patient characteristics is large, then conducting trials for patient types is simply impossible.

Ultimately, the goal of measuring the treatment effect is to make a better decision. Section 3 discusses the two contrasting approaches to evaluating decisions. As an example, consider data for the following scenario. Patient i with observed characteristics x_i seeks treatment from physician j, who then decides upon treatment choice, $d_i = 0$, *or* $d_i = 1$. The consequence is outcome, u_i^0 or u_i^1, depending upon the choice d_i. The conditional average treatment effect (CATE) is $\tau(x) = E\left\{u_i^1 - u_i^0 | x_i = x\right\}$. As Holland (1986) emphasises, the pair $\left\{u_i^0, u_i^1\right\}$ are potential outcomes only

[2] Angrist and Pischke (2010) on page 24 state that "This point has long been understood in medicine, where clinical evidence of therapeutic effectiveness has for centuries run ahead of the theoretical understanding of disease."

[3] This point is not new. See Ludwig et al. (2009).

- in practice we only observe $u_i^{d_i}$, and not $u_i^{d_i'}$, when $d_i' \neq d_i$. Notice that we can view randomized trials and perfect experts as two extreme ways to learn from data. RCTs are data sets where the decisions are by construction free of randomness, and hence with enough data we can construct estimates of the CATE $\tau(x)$, which in turn can be used to optimally treat a patient by seting $d_i = 1$ iff $\tau(x_i) \geq 0$.

In constrast, suppose we have a perfect decision maker who set $d_i = 1$ iff $u_i^1 \geq u^0$. Like the Roy model, since we only observed the optimal choice, without additional assumptions the counterfactual return is not observed, and hence the CATE cannot be estimated. However, the data is very informative. In fact, one goals of the literature on pattern recognition (Devroye et al. (1996)) is to take such data and build a decision function $d^*(x)$. In fact, as Devroye et al. (1996) discusses in Section 6.7, one needs less data to construct $d^*(x)$ from a perfect decision maker than to construction the CATE using regression techniques. In other words, if the goal is simply to get the best decisions, then having data with good decisions is more useful.

Section 4 discusses the human capital approach that combines both ideas. The starting points are the two contrasting views of experts (Kahneman and Klein (2009)). An expert is an individual who can make high quality decisions very quickly. For example, something as common place as driving requires the ability to process and react in real time to a complex stream of information. Even if one is not an "expert driver", driving requires an amazing combination of skills. In the context of medical decision making, the first step in our procedure is to suppose that physicians are experts, hence there is a positive relationship between their decision and whether or not the patient is better off getting treatment. Using machine learning 101 (the logistic regression - see Hastie et al. (2009)), we can use the full data set to determine the probability that a patient with characteristics x gets treatment, given by $\eta(x) = E\{d_i|x\}$.

The probability $\eta(x)$ is the familiar propensity score. However, the interpretation here is quite different than in the econometrics literature, where it has been controversial.[4] The difficulty with estimating the CATE when the feature space X is high dimensional is that it is not clear how to create groups within which the treatment effect is relatively constant. Here we are using experts to effect a dimension reduction that then allows one to apply the results from Rosenbaum and Rubin (1983). In Section 4 I show that one can provide a simple, decision theoretic model to justify this approach.

The second step entails estimating the CATE as a function of the propensity score. Here we are relying upon the second feature of expert decision making. Given that the acquisition of human capital is expensive, this implies that decision making is imperfect. Within the context of the simple model, the choice of action conditional upon the propensity score is assumed to be noisy. It is quite common to suppose that physician practice style is represented by a one dimensional fixed effect (e.g. Chandra and Staiger (2007)). In the context of this model, we characterize physician decision

[4]See Smith and Todd (2005) and the rejoiners.

making as two dimensional, where one dimension is the sensitivity of decision to the propensity score, which in turn can be interpreted as decision making skill.

In Section 5 I discuss two papers that use this approach to study the decision making skill of physicians treating heart attacks and assisting in childbirth. In that data we have physician identities, and hence we can directly test whether or not there is variation is decision making skill. In Currie and MacLeod (2017) we do indeed find that physicians who exhibit less sensitivity to patient conditions have worst outcomes on average, consistent with the hypothesis of poorer information. In Currie et al. (2016) we have a quite a different result. There we find that the CATE does *not* change sign with the propensity score - namely the evidence is consistent with the hypothesis that heart attack patients are always better off with the most invasive procedures. In that case variation in treatment is associated with non-medical characteristics of the patient.

The final section of the paper has some concluding remarks, and suggestions for future research.

2 The Rubin/Holland Model[5]

In this section I review the well known Rubin-Holland model outlined by Holland (1986), and explicitly link it to optimal decision making.[6] The question is how to use evidence from an experiment or observational data to make better decisions. I will reiterate the basic point in Holland (1986) that measuring a causal effect requires making some untestable assumptions. In practice these assumptions are typically implicit, rather than explicit, which in turn can lead to overly strong claims in some cases (see Deaton (2010)).

We begin with a universe of individuals whose characteristics are described by a compact set $X \subset \Re^n$. For example, this might be all persons in a country in the year 2000, or all individuals who had a fever last year. Individuals may also be firms or countries, though for the current discussion we can think of them as a collection of persons denoted by:

$$U = \{i \in P | x_i \in X\},$$

where x_i is the characteristic of individual i, and P denotes the universe of all possible individuals. Here I deviate slightly from Holland where the primitive is typically the set P. The reason is that the external validity of any experiment is defined by the set of persons for whom the results are valid. These individuals are typically not listed, but described by features such as race or where they live. Notice that this formulation includes as the special case in which each person is a unique point in X.

[5] Xuan Li did the background research on the effects of the psychiatric drugs. After the paper was accepted, we learned of the more comprehensive study by Cipriani et al. (2016) that comes to similar conclusions.

[6] See Imbens and Rubin (2011) for a comprehensive review of the approach and the historical background. See also Freedman (2006).

For each person i, we would like to know for each choice $d_i \in \{1, 0\}$, the set of *potential outcomes:*

$$\left\{ \left(x_i, u_i^1, u_i^0 \right) | i \in U \right\},$$

where u_i^1, u_i^0 are the outcomes for choices 1 and 0 respectively. These are potential outcomes because the choice is made at a given date, with payoffs realized in the future, and hence for each unit we can at best observe u_i^1 or u_i^0, but not both. I maintain throughout the *stable unit treatment value assumption (STUVA)* - the decision for unit $j \neq i$ does not affect the potential outcomes for unit i. The *average treatment effect (ATE)* of choice 1 is given by:

$$\tau^{ATE} = E\left\{ u_i^1 - u_i^0 | i \in U \right\}.$$

This is the parameter estimated with a randomized control trial (Imbens and Rubin (2011)). One procedure to measure ATE is as follows. Randomly select from U - the set of individuals that match the criteria in set X - $2n$ individuals, who are randomly assigned to group 1 - U_1 and group 0 - U_0. This generates data, $Data\left(n\right) = \left\{ x_i, u_i^{d_i} | i \in U_A \cup U_B \right\}$, where $d_i = 1$ if $i \in U_1$ and $d_i = 0$ if $i \in U_0$. The point here is that $Data(n)$ cannot contain both potential outcomes for the same unit, but it can be used to compute an estimate of average treatment effect:

$$\hat{\tau}^{ATE}\left(Data(n) \right) = \frac{1}{n} \left\{ \sum_{i \in U_1} u_i^1 - \sum_{i \in U_0} u_i^0 \right\}.$$

When the assignment is random ($x_i \perp\!\!\!\perp d_i$), then we have the well known result:

Proposition 1. *If units are randomly assigned to choices* 1 *and* 0*, and the stable unit treatment value assumption is satisfied, then the average treatment effect satisfies:*

$$\tau^{ATE} = E\left\{ \hat{\tau}^{ATE}\left(Data(n) \right) \right\} = lim_{n \to \infty} \hat{\tau}^{ATE}\left(Data\left(n\right) \right).$$

Proof. We follow Deaton (2010). First:

$$\begin{aligned} E\left\{ \hat{\tau}^{ATE}\left(Data(n) \right) \right\} &= \frac{1}{n} \left\{ \sum_{i \in U_1} E\{ u_i^1 | d_i = 1 \} - \sum_{i \in U_0} E\{ u_i^0 | d_i = 1 \} \right\}. \\ &= E\left\{ u_i^1 | d_i = 1 \right\} - E\left\{ u_i^0 | d_i = 0 \right\} \\ &= lim_{n \to \infty} \hat{\tau}^{ATE}\left(Data\left(n\right) \right) \end{aligned}$$

Next observe that:

$$\begin{aligned} E\left\{\hat{\tau}^{ATE}\left(Data(n)\right)\right\} &= E\left\{u_i^1|d_i=1\right\} - E\left\{u_i^0|d_i=0\right\}, \\ &= E\left\{u_i^1|d_i=1\right\} - E\left\{u_i^0|d_i=1\right\}, \\ &= E\left\{u_i^0|d_i=1\right\} - E\left\{u_i^0|d_i=0\right\}. \end{aligned}$$

Observe that by SUTVA and random assignment, we have that the final line is zero. Random assignment also implies that the expected value of a potential outcome (observed or not) is not affected by the assignment. Hence we have:

$$\begin{aligned} lim_{n\to\infty}\hat{\tau}^{ATE}\left(Data\left(n\right)\right) &= E\left\{u_i^1|d_i=1\right\} - E\left\{u_i^0|d_i=1\right\}, \\ &= E\left\{u_i^1 - u_i^0|d_i=1\right\}, \\ &= E\left\{u_i^1 - u_i^0|i\in U\right\}, \\ &= \tau^{ATE}. \end{aligned}$$

□

Though quite simple, this result nicely illustrates the power of RCTs - under the appropriate assumptions they allow for the measurement of the average treatment effect for a *population*. There is a large literature on constructing bounds to τ^{ATE} given finite data from an RCT. Our concern here is not with the implementation details for an RCT, but with the problem of making *decisions* using observational data.

The first condition, $\tau^{ATE} = E\left\{\hat{\tau}^{ATE}\left(Data(n)\right)\right\}$, is called the *ignorability condition.* It means that regardless of the sample size, the mean is an unbiased estimate of the treatment effect. However, this is no longer true for selected sub-samples, particularly sub-samples chosen as a function of x_i. The literature on estimating treatment effects has for the most part focused upon the problem of inferring τ^{ATE} as a function of different assignment mechanisms. In many cases, as both Deaton (2010) and Heckman (2010) observe, one may also be interested in the treatment effect for sub-populations of X.

For example, consider the problem of choosing a drug for the treatment of depression. In order for a company to sell a drug they have patented, it must go through trials with human subjects. Successful drugs provide a great deal of revenue to companies during the life of the patent, as we can see in Table 2. Thus they have a large financial incentive to have a successful trial and use the results of the trial to direct physicians on how to use a new drug.

We can view trials as have having three outcomes, $u_i \in \{V, 0, -L\}$, where $V > 0$ is to feel well, 0 is to be depressed, and $-L < 0$ is to commit suicide. The target populations are individuals who are currently depressed, denoted by X^D. The goal of treatment is to obtain the outcome $u_i = V$.

Table 1: Sales of SSRI drugs and mood stabilizers in the US

Drug type:	SSRI				Mood Stabilizer			
Drug name:	Lexapro (Forest Laboratories)		Zoloft (Pfizer)		Abilify (Otsuka Pharmaceutical)		Lamictal (GlaxoSmithKline)	
	Sales	Rank	Sales	Rank	Sales	Rank	Sales	Rank
2003	965,666	34	2,580,509	5	364,546	88	582,281	56
2004	1,551,230	17	2,622,801	5	747,400	47	780,614	43
2005	1,849,528	13	2,561,069	6	1,098,379	29	1,031,307	34
2006	2,098,794	10	1,772,599	15	1,417,106	24	1,326,844	26
2007	2,304,364	9	175,209	170	1,781,562	15	1,717,429	17
2008	2,412,048	11			2,371,795	12	1,539,101	19
2009	2,334,422	13			3,083,351	6	498,599	73
2010	2,483,391	12			3,514,265	6	326,331	101
2011	2,835,216	18			5,032,032	4		
2012					5,602,876	2		
2013					6,293,801	1		
Patent expiration	March 2012		June 2006		October 2014		Mid 2008	

Notes: Sales in the US in $000. Source: http://www.drugs.com/top200.html

The difficulty is that in order to get approval to use human subjects one cannot enroll patients into the study that are at high risk of suicide, but rather the subset of patients that are depressed, but not at risk of suicide:

$$\bar{X}^D = \left\{ x \in X^D | Pr\left[u_i = -L | x_i = x \right] \simeq 0 \right\}.$$

It is worth highlighting the fact that the drugs in Table 2 may elevate the risk of suicide for adolescents, but by construction these subjects are excluded from these trials. Yet, once approved, psychiatrists are free to prescribe these drugs as they wish, including prescribing them to adolescents (which is very common).

Second, one needs an instrument to measure the outcome of the trial. Since the trials are over relatively short periods, these outcome measures are at best proxies for the long term outcome (such as death by suicide). Such instruments are performance scores denoted by y_i. Again, one can only measure the outcome of the chosen treatment and not both potential outcomes. The *extended Rubin/Holland model* is concerned with measuring both the performance scores and the outcomes:

$$\left\{ x_i, \left\{ y_i^1, y_i^0 \right\}, \left\{ u_i^1, u_i^0 \right\} \right\}_{i \in U}.$$

In the case of depression, drug researchers use the Montgomery-Asberg Depression Rating Scale (MADRS), Hamilton Rating Scale for Depression (HAMD), or Children's Depression Rating Scale-Revised (CDRS-R) to produce a score before and after treatment, y_i and $y_i^{d_i}$.[7]

We then set:

$$\begin{aligned} \Delta Score_{treat} &= y_i^1 - y_i, \\ \Delta Score_{placebo} &= y_i^0 - y_i. \end{aligned}$$

The average treatment effect is then defined by:

$$Relative\ Score\ Reduction\ (RSR) = \frac{\Delta \hat{Score}_{treat} - \Delta \hat{Score}_{placebo}}{\Delta \hat{Score}_{placebo}},$$

where the hat refers to the population means. The results from a number of studies looking at Lexapro and Zoloft are reported in Tables 3 and 4.[8] The average treatment effect is reported in the column RSR. The RRR column is computed in the same way using the fraction of individuals whose depression rate is reduced by 40%-60%.

The decision to prescribe a drug is based upon the trials such as the ones in Tables 2 and 3.

[7] See Cusin et al. (2010)

[8] Studies looking at Lexapro are: Lepola et al. (2003), Wade et al. (2002), Burke et al. (2002), Pigott et al. (2007), Azorin et al. (2003), Bech et al. (2004), Ninan et al. (2003), Llorca et al. (2005), Ventura et al. (2006), Findling et al. (2013), Emslie et al. (2009), Wagner et al. (2006).Studies of Zoloft include Ventura et al. (2006), Stahl (2000), Fabre et al. (1995), Olie et al. (1997), Schneider et al. (2003), Wagner et al. (2003), Donnelly et al. (2006), March et al. (1998).

Table 2: Results from Randomized Control Trials for Lexapro (Escitlopram)

Study	**Citations**	**# treatments**	**# placebo**	**Age**	**Dosage (mg)**	**Duration**	**RST (p-value)**	**RRR (p-value)**
Lepola et al.[2003]	242	155	154	18-64	10 or 20	8 weeks	0.24 (0.002)	0.32 (0.06)
Wade et al. [2002]	228	191	189	18-64	10	8 weeks	0.20 (0.002)	0.31 (0.05)
Burke et al. [2002]	218	118	119	18-64	10	8 weeks	0.36 (0.002)	
		123			20		0.47 (0.002)	
Pigott et al.[2007]	67	274	137	18-75	10	8 months	0.35 (0.03)	
Azorin et al.(2004)	28	169	166	18-64	20	8 weeks	0.39	0.47 (0.05)
Bech et al.[2004]	67	118	119	18-64	10	8 weeks	0.22	
		123					0.3	
Ninan et al.[2003]	3	143	119	18-64	20	8 weeks	0.37	
Llorca et al. [2005]	93	163	166	18-64	10	8 weeks	0.37	0.43 (0.05)
Ventura et al.[2007]	51	78	79	18-80	10	8 weeks	0.36	0.44 (0.07)
Findling et al.[2013]	0	155	157	12-17	10 or 20	24 weeks	0.23 (0.001)	0.35 (0.05)
Emslie et al.[2009]	77	155	157	12-17	10 or 20	8 weeks	0.17	
Wagner et al.[2006]	136	133	131	6-17	10 or 20	8 weeks	0.08 (0.31)	

Notes: There are many RCTs which assign subjects to different treatment groups without placebo control. Here I include those RCTs in which an explicit placebo group is assigned. Google scholar citations up till Feb 20, 2015 are reported. RSR stands for relative score reduction and RRR stands for relative response rate.

Table 3: Results from Randomized Control Trials for Zoloft (Sertraline)

Study	**Citations**	**# treatments**	**# placebo**	**Age**	**Dosage (mg)**	**Duration**	**RST (p-value)**	**RRR (p-value)**
Ventura et al.[2007]	51	85	79	18-80	50-200	8 weeks	0.27	0.34 (0.07)
Stahl et al.[2000]	190	108	108	18-75	50-150	8 weeks	0.27	
						24 weeks	0.41	
Fabre et al.[1995]	156	95	91	18-75	50	6 weeks	0.29	
		92			100		0.32	
		91			200		0.54	
Olie et al.[1996]	19	129	129	18-70	50-200	6 weeks	0.48	0.57 (0.06)
Schneider et al.[2014]	143	371	376	>=60	50-100	8 weeks	0.12	
Wagner et al.[2003]	136	189	187	6-17	50-200	10 weeks	0.17	0.17 (0.05)
Donnelly et al.[2006]	15	103	106	12-17	100	10 weeks	0.18	0.28 (0.07)
March et al.[1998]	425	92	95	6-17	200	4 weeks	1	0.43 (0.07)

Notes: There are many RCTs which assign subjects to different treatment groups without placebo control. Here I include those RCTs in which an explicit placebo group is assigned. Google scholar citations up till Feb 20, 2015 are reported. RSR stands for relative score reduction and RRR stands for relative response rate.

In general the point estimates are all positive. This leads practitioners to prescribe the medication because they believe that credible RCTs suggest that they work. Yet, as Ludwig et al. (2009) observe, these results lack external validity because individuals at risk of suicide must, for ethical reasons, be excluded from the studies.[9]

Moreover, the outcome of these trials is an index whose value does not have an obvious economic interpretation. That is to say, there is no obvious weighting rule that, for example, includes the loss in value due to completed suicides; hence the average treatment effect may not reflect the optimal choice. We also know that SSRIs may have significant side effects, and hence any treatment effect should include values associated with illness caused by the drug.[10]

The American Psychiatric Association looked at the question of how treatment affects suicide rates. The results for different age groups are shown in Table 5. As one can see, the success of treatment for younger patients is definitely mixed. In particular, for younger patients these drugs may increase the risk of suicide, and they are now packaged with "black box" warnings to this effect. Given that by age 25 suicide has already claimed individuals, the positive effect at that age may be due in part to the selection effect of suicide!

Table 4: Suicidality from a Meta-study of RCTs by American Psychiatric Association

Age Range	Drug-Placebo Difference in Cases of Suicidality /1000 Patients
<18	14 additional cases
18-24	5 additional cases
25-64	1 additional case
>=65	6 fewer cases

Notes: Results are from RCTs on all andidepressants for patients with MDD, Obsesive Compulsive Disorder (OCD), or other psychiatic disorders.

Currently, it is very difficult to determine whether a patient with certain characteristics $x \in X$ will benefit from treatment. The question then is how to use these results to guide decision making in practice. For simplicitly, suppose that individuals are one of three types. For type A, given by $x \in X^A$ treatment with the drug cures the depression with certainty, resulting in the payoff V. Similarly, for a type B person, $x \in X^B$, treatment has no effect, while for type C, $x \in X^C$, the result is suicide and a lost of $-L$. Let $p^t, t \in \{A, B, C\}$ be the population probabilities for each type. Under the hypothesis that the physicians cannot tell which type they face, then the appropriate criteria for treatment is the average treatment effect:

$$\tau^{ATE} = p^A V - p^B L.$$

[9]Ludwig et al. (2009) use observational data and the fact that variation in the way the drugs are priced and distributed affects the level of SSRI usage. Using population level measures of suicide rates, they find that an increase in the class of selective serotonin re-uptake inhibitors of 1 pill per capita (12% of 2000 sales levels) reduces suicide by 5%.

[10]For the FDA warnings on Zoloft and Lexapro go to `http://www.fda.gov/Drugs/DrugSafety` and search for the drug specific information.

This example illustrates the challenge one faces when using an RCT to evaluate treatment. First, neither the benefit (V) nor the cost (L) from the potential outcomes can be directly measured. Hence, techniques such as those in Hirano et al. (2003), used to obtain efficient estimates of the average treatment effect cannot be used. Second, there is the obvious sample selection problem because individuals are restricted to have characteristics in $X^A \cup X^B$, those not at immediate risk from suicide.

An alternative approach focuses upon evaluating *decision rules* rather than the treatment effect. Specifically, can we identify the set of characteristics X^+ such that $\tau(x_i) > 0, \forall x_i \in X^+$. This in turn determines a decision rule that improves upon a rule based upon the ATE by allowing choice to vary with observed characteristics. We now turn to this question.

3 The Evaluation of Decision Rules

The evaluation of drugs for the treatment of depression illustrates some of the challenges one faces when using randomized trials to address a substantive issue. In addition to the *fundamental problem of causal inference* (Holland (1986)), due to the impossibility of observing both potential outcomes for the same unit, it is typically also the case that one cannot directly measure the outcome of interest. For example, in the case of depression, one only observes a proxy for the person's mental state. In terms of policy it is not obvious how to agregate such measures over a large population for purposes of providing general therauptic advice, such as the recommendation of an SSRI as the first drug to try for treatment.

An alternative approach would be to focus upon *decision rules* rather than the *treatment effect.* In this section I briefly discuss the evaluation of decision rules and how they compare to measures of the treatment effect. If we expect treatment have the same sign for the full population, say we want to know the average effect of a vaccine that will be delivered to the whole population, then it makes sense to use evidence from a sample of the whole population to obtain a more precise estimate of the effect (as in Hirano et al. (2003)). Heckman (2010) notes in passing one may also be interested in the *voting criteria.*[11] Under this rule we ask what *fraction* of the population would be better off from treatment. He mentions that this approach is used in political economy, and does not discuss it further. It turns out that this is also the approach used in the pattern recognition and machine learning literatures to evaluate the quality of the decision rules[12] Moreover, as Devroye et al. (1996) observe (sec 6.7), measuring decisions is easier than measuring treatment effects.

More precisely, given a unit $i \in U$, we can define two random variables that are unobserved, but can be used to define the performance of a decision rule. The realized treatment effect:

$$\tau_i = u_i^1 - u_i^0,$$

[11] Ibid, page 364.

[12] See Devroye et al. (1996) and Hastie et al. (2009).

and the best treatment choice:

$$d_i = \begin{cases} 1, & if\ \tau_i \geq 0, \\ 0, & if\ not. \end{cases}$$

Neither of these variables can be directly observed at the time choice is made. What we have are the observe characteristics of the individual, x_i, from which we can define the two parameters that are potentially estimable from data. The first the conditional average treatment effect:

$$\tau(x) = E\{\tau_i | i \in U, x_i = x\}, \tag{1}$$

and the probability that treatment is effective:

$$\eta(x) = E\{d_i | i \in U, x_i = x\}. \tag{2}$$

Ultimately, given that the characteristics of the unit i, conditional upon x_i, are observed before treatment, then we are interested in using data, either from an RCT or observational data, to choose a decision function:

$$d : X \to \{0, 1\}.$$

In the learning literature the norm is to evaluate decision functions using a loss relative to the best that can be obtained. There are two criteria one can use. The first, is the "economic" critera that supposes that the treatment effect is measured with transferable utility. In that case the *welfare loss* of a decision function is measured by:

$$WL(d) = E\{max\{\tau_i, -\tau_i\} | i \in U\} - \int_x \tau(x)(2d(x) - 1)\, d\mu(x). \tag{3}$$

The welfare lost is the diference between the maximum welfare if one chooses the most effect treatment for each individual, less the conditional treatment effect for each $x \in X$ determined by the decision rule. Where $\mu(x)$ is the distribution over characteristics. Clearly, $WL(d) \geq 0$ for all decision rules. The second criteria is the *Bayes Risk* defined by:

$$L(d) = Pr\{d(x_i) \neq d_i | i \in U\} \tag{4}$$

It measures the frequency with which a decision rule varies from the best rule, as opposed to a rule that takes into account the implicit cost of deviating from the optimal choice.

Associated with each rule are natual optimal decision rules. For the welfare loss we have:

Proposition 2. *For every measureable decision rule $d(.)$ we have $WL(d) \geq WL\left(d^{cate}\right)$ where:*

$$d^{cate}(x) = \begin{cases} 1, & if\ \tau(x) \geq 0, \\ 0, & if\ not. \end{cases}$$

The result follows immediately from an inspection of (3). Thus, if we are able to estimate the CATE $\tau(x)$, then a decision rule based upon this will provide the lowest expected loss relative to the theoretical maximum. In particular, if the sign of $\tau(x)$ changes over the set X, then the optimal rule should vary with x, and decision making based solely upon the average treatment effect cannot be optimal. In the case of the Bayes risk criteria we have:

Proposition 3. *For every measureable decision rule $d(.)$ we have $L(d) \geq L\left(d^B\right)$, where d^B is the optimal Bayes rule defined by:*

$$d^B(x) = \begin{cases} 1, & if\ \eta(x) \geq 1/2, \\ 0, & if\ not. \end{cases}$$

This result follows from Theorem 2.2 in Devroye et al. (1996). In this case, if the probability that treatment 1 is optimal is greater than $1/2$, then the optimal Bayes rule is to choose 1. This exactly Heckman (2010)'s voting rule. One chooses the decision that is more frequently correct, or,stated another way, is the majority of equally probable events select 1. There are some cases in which the criteria lead to the same choice. The first of these is when conditional upon x there is always an optimal choice:

Proposition 4. *Suppose $L\left(d^B\right) = 0$ and τ_i is bounded then $WL\left(d^B\right) = 0$ and the optimal CATE rule and Bayes differ at most on a set of measure zero.*

If $L\left(d^B\right) = 0$ then this implies that almost everywhere $\eta(x) \in \{0, 1\}$, and hence there is best decision for almost every $x \in X$. From this it follows that $WL\left(d^B\right) = 0$.

This result is useful because when we are in a situation where there is clearly a correct choice for each $x \in X$, then the size of the treatment effect is not relevant for setting the decision rule, only the sign is relevant. For example, this provides some guidance regarding the use of proxy variables in an RCT. If a drug helps relieves depression if and only if the patient has a better Montgomery-Asberg Depression Rating (MADRS) or Hamilton Rating (HAMD), then the results from an RCT for SSRIs can be used in clinical practice to recommend treatment, even though the value of treatment is deficult to measure.

In many cases there is no clear, unambiguously correct choice. This can occur when there are unobserved factors that affect the CATE, but they are not contained in the vector of observed person characteristics, x_i. Even so, there is a case in which the optimal rule based upon the treatment effect and the Bayes optimal rule imply the same optimal choice.. Suppose that the distribution of

τ_i is symmetric around it's expected value $\tau(x)$ for all $x_i = x \in X$, and $Pr\{\tau_i = \tau(x_i)\} = 0$ (there is no mass at $\tau(x)$). Then $Pr\{\tau_i < \tau(x)\} = Pr\{\tau_i > \tau(x)\} = 1/2$, and we have that $\tau(x) \geq 0$ if and only if $\eta(x) \geq 1/2$. Thus:

Proposition 5. *Suppose that the treatment effect τ_i is symmetrically distributed around $\tau(x)$, with no mass at $\tau(x)$, then the optimal CATE rule (d^{CATE}) and the optimal Bayes rule (d^B) are the same almost everywhere.*

Finally, the two approaches represent contrasting approaches on how to learn from data. Notice that while we can never directly observe the treatment effect τ_i, we *can* observe decisions made by agents, and the consequence of these decsions, either $u_i^{d_i}$ or $y_i^{d_i}$. Randomized control trials represent one extreme, where the decision d_i is explicitly randomized so that we sufficient data we can estimate $\tau(x)$ from observations of outcomes. In that case the decision rule itself contains no information.

In contrast, consider the other extreme case in which the optimal Bayes risk is zero - $L\left(d^B\right) = 0$, and we have data from perfect expert decision makers who choose $d_i = 1$ iff $\tau(x) \geq 0$. In that case we never observe the counterfactual inefficient choice, and hence have *no* information concerning the treatment effect. However, we are in a situation in which we can learn the decision rule. In this case, as Macleod (2016) discusses, with enough observations it is possible to estimate the optimal decision rule for all $x \in X$, even though measuring the treatment effect is impossible.

The traditional approach in empirical labor economics is to view any correlation between the treatment effect and choice as creating a threat to identification (Angrist and Krueger (1999)). It is worth highlighting the point that the large literature on pattern recognition and machine learning takes exactly the opposite view. The more tightly connected choice is to the optimal treatment, the lower the Bayes risk, which in turn improves the ability of algorithms to learn the best choices from training data. In the next section I discuss some recent work that combines these viewpoints and illustrates how we can use a mixed approach to learning on how to improve observed decision making in medicine.

4 The Human Capital Approach to Inference

This section outlines what I call the *human capital* approach to inference. The goal is to provide a way to lever expert knowledge, or human capital, to estimate a version of the CATE that in turn can lead to improvements in decision making. The standard approach to identify CATE is knowledge of the environment that allows one to put some structure upon the assignment to treatment groups. The instrumental variables approach, such as Angrist et al. (1996), assumes that there is some shock in the environment that creates a random assignment. Vytlacil (2002) and Heckman (2010) observe that the Roy model can be interpreted as a valid estimate of the returns to changing sectors by viewing moving costs between sectors as an exogenous shock that is

independent of the treatment effect. Athey and Imbens (2015) discuss the use of machine learning techniques to measure the CATE, but still rely upon the exogeneity of the treatment effect (as in Theorem 1).

Here I begin with an environment with many heterogeneous units, and at least two (but not an infinite number of) agents who carry out the assignment to treatments. The precise context we have in mind is a physician $j \in J$ treating patient $i \in U_j$ with condition x_i. The set U_j indexes the patients for physician j, with the feature that $U_j \cap U_{j'} = \emptyset$ whenever $j \neq j'$ and $\cup_{j \in J} U_j = U$. Matters are much easier if we suppose that the distribution of x_i for $i \in U_j$ is given by μ for all $j \in J$. This is a strong assumption, and we defer discussion of it to the end. The job of the physician j is to choose treatment $d_{ij} \in \{0, 1\}$ as a function of the observable conditions of patient i, given by $x_i \in X$, where X is a finite set.[13] In the spirit of the SUTVA, I assume that physicians treat "in a bubble."[14] Namely, their treatment decisions are fixed when they leave medical school. For the current discussion - Epstein and Nicholson (2009) provide some direct evidence in support of this assumption.

The problem is made more complex by that fact that the number of possible conditions represented in the set X is potentially large. The purpose of medical school is to teach students the best way to treat patients as a function of $x \in X$, so that they make decisions that are close to optimal, which we suppose is given by $d^*(x)$.

When we say that this decision making ability is *human capital,* this has two implications. The first is that it is expensive to acquire. As I point out in Macleod (2016), this implies that decision making is *imperfect,* but increasing with experience and the innate ability of the individuals. Even highly skilled individuals make mistakes. These errors create random assignment from which we can determine the treatment effect. The second implication is that even though physicians make errors, they are not random. Millions of individuals are treated by physicians each year with the expectation that treatment by a physician is better than the alternative.

This implies that the allocation to a treatment is non-random. We can exploit this fact and use a basic machine learning algorithm to organize the data before attempting to exploit error to measure the CATE. More precisely, let us suppose that Agent $j \in J$ has an *unbiased* noisy observation of the CATE (1):

$$\tau_{ij}(x) = \tau(x) - \epsilon_{ij}, \tag{5}$$

where $\epsilon_{ij} \sim N\left(0, \sigma_j^2\right)$, where $\sigma_j^2 > 0$ is constant for each doctor. A smaller variance σ_j^2 corresponds to more diagnostic skill. I am assuming that the treatment effect is on a log scale, so that τ and y_i take values from $(-\infty, \infty)$. If training were perfect and homogeneous, then we would suppose that $\sigma_j^2 \simeq 0$. We begin with the hypothesis that the quality of decision making among the $j \in J$

[13] Not only does this simplify the analysis, but is also true in practice since any information reporting system by construction has only a finite number of possible x variables.

[14] This is a direct quote from a physician, who said that after medical school his decision making was independent of other physicians' decisions.

Agents varies with the variance σ_j^2. There is quite a bit of evidence that this is the case. In the case of physicians, there is a large amount of variation in practice styles that cannot be explained by the condition of the patient, an observation that is often used to explain the high cost of health care in the U.S., along with the under-provision of care in other cases (Song et al. (2010)).

Let us suppose that we have a data set given by:

$$\begin{aligned} Data &= \left\{ \left\{ x_i, u_i^{d_{ij}} | i \in U_j \right\} | j \in J \right\}, \\ &= \left\{ Data_j | j \in J \right\}. \end{aligned} \tag{6}$$

With this data we would like to answer two questions. First, do physicians vary in quality of decision making? Second, what are the features of the better doctors? In particular, we would like to offer specific guidance on how their decisions might change to improve outcomes. We begin with the pattern recognition or matching learning approach to thinking about a decision. Consider physician j. Their job is to divide patients into two groups, X_j^1 and X_j^0, and then carry out the decision:

$$d_j(x_i) = \begin{cases} 1, & x_i \in X_j^1, \\ 0, & x_i \in X_j^0. \end{cases}$$

What one learns in medical school are patient conditions that determine the sets X_j^1 and X_j^0 - the problem of pattern recognition is to take the observed data to reconstruct these sets. The assumption that a doctor observes a noisy signal of the treatment effect dramatically complicates the problem. Given the learning process (5), then the set of conditions where $d_j(x) = 1$ is given by conditions $x \in X^1$ such that the physician believes the best course of action is to treat. This set includes x if there is a chance that $\tau_{ij}(x) > 0$. Since ϵ_{ij} is Normally distributed, then it's support is unbounded and we have:

$$\begin{aligned} X_j^1 &= \{ x \in X | for\ some\ i, \tau_{ij}(x) = \tau(x) - \epsilon_{ij} \geq 0 \}, \\ &= X\ with\ prob\ 1,\ as\ \#U \to \infty. \end{aligned}$$

In other words with a noisy signal there is always a chance a physician might recommends $d_i = 1$ and $X_i^1 = X_i^0 = X$! Hence, for each $x \in X$ the probability of treatment is in $(0, 1)$.

The human capital approach to inference used here relies on a few assumptions. First let us suppose that for a randomly selected individual the probability of using physician j is ρ_j. Suppose that for this individual the CATE is τ, then the probability of getting treatment 1 is:

$$\begin{aligned} e(\tau) &= Pr[d_i = 1 | \tau], \\ &= \sum_{j \in J} \rho_j F\left(\frac{\tau}{\sigma_j} \right). \end{aligned} \tag{7}$$

The assumption that decision making is imperfect implies that $\sigma_j > 0$, and hence:

$$e'(\tau) = \sum_{j \in J} \rho_j f\left(\frac{\tau}{\sigma_j}\right) / \sigma_j > 0. \tag{8}$$

This implies a 1-to-1 relationship between the probability of treatment and the treatment effect τ. This function is the familiar *propensity score.* Since the score is strictly increasing with τ, then it becomes a *balancing score* in the sense of Rosenbaum and Rubin (1983), because conditioning upon e allows for a consistent estimation of $\tau(x)$. The first step is to construct the population propensity score as a function of the data:

$$\eta(x) = E[d_i | x_i = x].$$

This is connected to the propensity score via $\eta(x) = e(\tau(x))$. We have:

Proposition 6. *Suppose that the SUTVA is satisfied, $e'(\tau) > 0$ for all $\tau \in \Re$, $\eta(x) = E\{d_i | x_i = x\}$ and $\bar{\eta} = \eta(\bar{x})$, then if:*

$$\bar{\tau} = E\left\{u_i^1 | d_i = 1, \eta(x_i) = \bar{\eta}\right\} - E\left\{u_i^0 | d_i = 0, \eta(x_i) = \bar{\eta}\right\},$$

it follows that $\eta(x_i) = e(\bar{\tau})$ for all $x_i \in \{x | \eta(x) = \bar{\eta}\}$ and $\bar{\tau} = \tau(x_i)$, the CATE at x_i.

Proof. Under the SUTVA the propensity score is a balancing score, and from theorem 4, Rosenbaum and Rubin (1983), $\bar{\tau}$ is the CATE at $e(\bar{\tau})$. The fact that $e' > 0$ implies that it is unique, and hence $CATE = \bar{\tau}$. □

We are making two key assumptions. First, the probability of treatment increases as a function of τ for each physician, but it is not perfectly correlated. This is the essence of the human capital approach - we suppose that doctors on average respond correctly to patient condition. Second we have assumed the allocation of patients to doctors is independent of the treatment effect. This is not strictly necessary since $e'(\tau)$ is strictly positive. All that is necessary is that the proportions do not change too quickly with τ.

We can perform some additional robustness checks. In this setup we are assuming that the physicians are making errors conditional upon the information they have in x_i. If that is true, then if we compare two physicians, and $\sigma_j^2 > \sigma_{j'}^2$, when j' is a better doctor, her propensity score rises more quickly. With sufficient data we estimate $\eta_j(x) = \eta\left(x, \sigma_j^2\right) \equiv F\left(\frac{\tau(x)}{\sigma_j}\right)$, the Agent's probability of treatment, by restricting the sample to a single agent j. The expected performance

of Agent j is given by:

$$\begin{aligned} Q_j\left(\sigma_j^2\right) &= \int_{x\in X} \eta_j(x)\,\tau(x) - (1-\eta_j(x))\,\tau(x)\,d\mu(x) \\ &= \int_{x\in X} \tau(x)\,(1-2\eta_j(x))\,d\mu(x) \end{aligned}$$

A simple computation implies:

Proposition 7. *The Agent-specific propensity scores and performance satisfy:*

$$\begin{aligned} \frac{\partial \eta_j(x)}{\partial \sigma_j} &< 0, iff\ \tau(x) > 0, \\ \frac{\partial Q_j\left(\sigma_j^2\right)}{\partial \sigma_j} &< 0. \end{aligned}$$

These results follow immediately from differentiating the respective expressions. Since $\eta_j(x) = 1/2$ iff $\tau(x) = 0$, this implies that for $\eta_j(x) > 1/2$, increasing the quality of information (lower σ_j) results in a higher probability of treatment, with the opposite occurring for $\eta_j(x) < 1/2$. Thus the quality of information has an *ambiguous* effect upon choice. In contrast, increasing the the quality of information (lower σ_j) always increases total performance.

What we have done is provide some structure to the well known propensity score model that allows us to interpret a propensity score as a *decision* rather than a self-selected treatment. The result relies upon two features of human capital:

1. Agents $j \in J$ are skilled in the sense that the propensity score to treat should rise with the treatment effect.

2. Human capital is expensive to acquire, and hence decision making is imperfect, which in turn implies that conditional upon the propensity score we are observing both potential outcomes.

5 Example: Medical Decision Making

A common approach to the estimating a treatment effect involves the creation of well defined groups, within which assignment to treatment and control are independent of individual characteristics. In contrast, here it is assumed that agents are making decisions to treat based upon their own perception of the efficacy of treatment. If their decisions are error free, then we would observe a great deal of homogeneity in their decisions. Moreover, if choice is perfect, then it is impossible to estimate the treatment effect because we only observe the optimal choice, not the counter-factual. However, the fact that experts do make mistakes creates heterogeneity in treatment that we can use to estimate the treatment effect. In this section I discuss two papers that apply these ideas to physician decision making.

In both cases it is assumed the physician decides whether or not to treat a patient with an invasive procedure. In the case of heart attack patients this is angioplasty or catheterization, while in the case of birth it is the choice between a natural delivery or a C-section. We begin by estimating $\eta(x)$, the population level probability that a patient with characteristics x_i is treated intensively.[15] This can be viewed as a classic problem in machine learning. Given $Data$, can we predict what will happen to a patient with characteristics x_i? As it turns out, the standard logit model is a very good machine learning model:[16]

$$\hat{\eta}(x) = Pr[d_i = 1 | x_i = x] = F(\Gamma x), \tag{9}$$

where $d_i = 1$ indicates an invasive procedure, F is the logit function, and Γ is a vector of parameter estimates. We then divide patients into two groups - high and low appropriateness for an invasive procedure:

$$\begin{aligned} U^H &= \left\{ i \in U | \hat{\eta}(x_i) \geq p^H \right\}, \\ U^L &= \left\{ i \in U | \hat{\eta}(x_i) \leq p^L \right\}, \end{aligned}$$

where p^H and p^L are chosen to create approximately three groups of individuals of equal size. In general, the index $\hat{\eta}(x)$ provides a way to rank patients along one dimension based upon how they are treated in the market.

The next issue is whether or not there is variation in the decisions made by the doctors. We estimate this by defining an index for patient condition $s(x) \in (-\infty, \infty)$ by:

$$\hat{\eta}(x) = F(s(x)).$$

For each physician we estimate the individual behavior for $i \in U_j$ via:

$$\hat{\eta}_j(x) = Pr[d_i = 1 | x_i = x] = F(\alpha_{jt} + \beta_{jt} s(x)), \tag{10}$$

where $\{\alpha_{jt}, \beta_{jt}\}$ is a physician's *practice style* at date t. If a physician behaved exactly the same as his or her colleagues, then the estimated values should not be significantly different from $\{0, 1\}$.

In order to evaluate the effect of practice style upon the patient we construct a measure of

[15] In the case of a heart attack patient, an invasive procedure is either angioplasty or catheterization (ICD codes 00.66, 36.0.., 37.22 or 37.23). For delivery of a child, a C-section is the invasive procedure.

[16] See Chapter 4, Hastie et al. (2009).

performance using observed outcomes for the each patient in the high and low categories:

$$\hat{u}_j^H = \frac{1}{n_j^H} \sum_{i \in U_j \cap U^H} u_i, \tag{11}$$

$$\hat{u}_j^L = \frac{1}{n_j^L} \sum_{i \in U_j \cap U^L} u_i, \tag{12}$$

where $n_j^l = |U_j \cap U^l|$ is the number of patients served by physician j in population $U^l, l \in \{L, H\}$. We then ask, do these measures vary systematically with physician practice style? Notice that an increase in α_j leads to more invasive procedures for *all* patients, while an increase in β_j leads to fewer invasive procedures for low risk patients and more invasive procedures for high risk patients. Let us now turn to the two applications.

5.1 Heart Attack Treatment

Currie et al. (2016) use hospital discharge data from all heart attack patients in Florida from 1994 until 2014. The question we ask is whether or not there is variation in physician decision making quality, and whether or not this is related to outcomes. We restrict the sample to heart attack patients who arrive at a hospital through the emergency room (ER) and are treated by a cardiologist. The result is a sample with 658,553 patients (U) treated by 2,929 cardiologists (J) at 149 hospitals. The set of patient characteristics (X) is listed in the first column of Table 5.

The index (9) is estimated using the data from teaching hospitals. This helps ensure that the index is based upon a group of skilled physicians. The patients for whom an invasive procedure is appropriate (U^H) are those with $\hat{\eta}(x_i) \geq .66$, while the low appropriateness patients (U^L) are those with $\hat{\eta}(x_i) \leq .34$, The mean values of x_{ik} for each group are listed in columns 3 and 4 of Table 5.

Next, for each physician $j \in J$, equation (10) is estimated. The first question we address is whether or not there is evidence that providers deviate significantly from the behavior of physicians in accredited hospitals. These results are presented in Table 6. We can see that there is significant deviation from the mean behavior in the market. About 13% of the physicians are less sensitive to patient conditions than the market mean, while 2% are more sensitive. The variation in the fixed effect is greater, with about 22% of the sample with a propensity to treat invasively regardless of the patient condition.

From these results we learn that there is no a consensus on how to treat these patients. This variation implies that by comparing the outcomes between physicians $j \in J$ we can learn what treatment styles are more effective because patient with similar characteristics are receiving different treatments. Table 7 presents the results from how variation in practice affects various outcomes for high and low appropriateness patients (versions of equations [11] and [12]). What is interesting is that more aggressive physicians get better outcomes. Also, low responsiveness physicians

get worse outcomes for the high appropriateness patients, while having better outcomes for low appropriateness patients.

Taken together, these results suggest that when judged from a purely medical point of view, a more aggressive treatment of heart attack patients leads to better outcomes. In general we find that U.S. trained physicians are less responsive and more aggressive, consistent with getting better medical outcomes. What is interesting is that physicians from top U.S. schools, while more aggressive, are also more responsive. As one can see from Table 5, one of the most important factors signaling aggressiveness is the age of the patient. Thus, it would seem that even though invasive procedures improve medical outcomes, for some patients, particularly older patients, some physicians are choosing to be less aggressive. This is consistent with them taking into account factors other than the treatment effect of an intervention.

5.2 Caesarean Sections

There is a great deal of concern that C-section rates at American hospitals are too high. In order to help mothers make better decisions, Consumer Reports (2015) provides advice on hospital choice, and recommends low C-section hospitals. Implicitly, they are making two assumptions. The first is that doctors at low-C-section hospitals have uniformly low C-section rates. However, while it is mechanically true that choosing a hospital with a low C-section rate results in a lower expected rate for the mother, Epstein and Nicholson (2009) find little relationship on C-section rates between physicians at the same hospital.

Second, the C-section rate recommendations that are used to evaluate physicians and hospitals assume that it is for a low risk pregnancy. Implicitly it is assumed that physicians will perform a C-section whenever it is medically necessary. Two questions remain. First, how should a mother decide if she is low risk or not? Normally, it is the job of the physician to do this, not the mother. Second, after a physician has been chosen and a preliminary evaluation has been carried out, there is the issue of the quality of decision making in real time during the labor and delivery process.

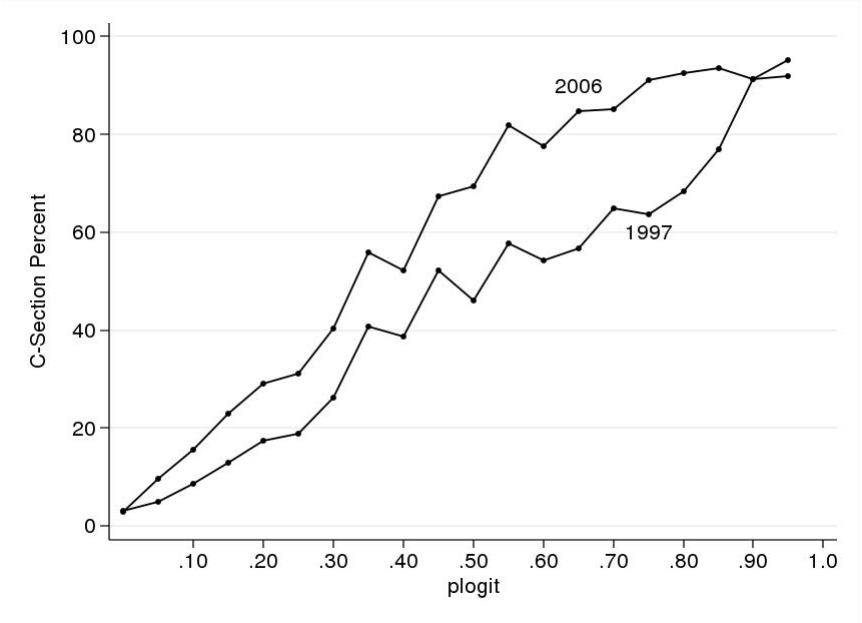
We already have strong evidence that physicians respond to financial incentives (Chandra et al. (2012)). In the specific case of C-sections, Currie and MacLeod (2008) find that obstetricians are responsive to changes in medical liability - in particular, when legal liability increases, obstetricians reduce their C-section rates for the marginal cases. This result is consistent with Johnson and Rehavi (2016) who find that when the mother is a physician, then she has a lower C-section rate and gets a better outcome. Yet, as Chandra et al. (2012) observe, incentives alone cannot explain all variation in practice style. Holding incentives fixed, a natural question is the extent to which there is variation in the way physicians make decisions, and does this variation lead to variation in outcomes?

Currie and MacLeod (2017) address this question using the human capital approach described above. They explore the quality of decision making using information from 1.1 million births

in New Jersey from 1997 to 2004. We are able to match these births to 71 hospitals and 5,273 birth attendants. Since only physicians carry out C-sections we remove the 603 midwives from the sample. For each delivery we have a rich set of X measures. These are listed in Table 8, along with the estimated coefficients for equation (9). We run the model for the full sample, as well as a sample of "good physicians" - those in the bottom 25th percentile of having any adverse outcomes. One can see that the rankings are very similar, with a correlation of .99.

What this ranking does is show that physicians rank $x_i \in X$ from different patients in the same way. We also know there has been a secular increase in C-section rates over time. The relationship between our index and the observed C-section rate is illustrated in Figure 1. We can see that there is a strongly positive correlation between our measure of risk of C-section with observed C-section rates. Also, the Figure documents the upward shift in C-section rates for all mothers, with the largest increase occurring in the 0.5 to 0.9 region. Given the changes over time, we allow estimated physician practice style to vary with time.

Figure 1: Shifts in Probability of a C-Section Over Time



Source: Figure 1, Currie and MacLeod (2017).

The next question is whether physicians vary systematically in the way they treat patients. In Currie and MacLeod (2017) we provide a formal model of physician decisions that provides a structural interpretation of equation (10). Specifically, physicians who are better at diagnosis have a higher β_{jt}. This is the case under the hypothesis that the index we construct accurately ranks patients, and that physicians make errors in their evaluation of patient condition. We will be able to check this hypothesis by seeing if variation in β_{jt} is associated with variation in outcomes, as

predicted by Proposition 4. An alternative hypothesis is that the physicians have better information than we have as outside observers. In that case we would expect the reverse – an increase in β_{jt} implies less private information, and hence worse outcomes. As we shall see, the data rejects this alternative hypothesis.

In addition, we measure procedural skill by calculating the rate of any bad outcomes among very low-risk births and the rate of bad outcomes among high-risk births for each doctor, and then take the difference between them. Taking the difference in the incidence of bad outcomes between these two groups is suggested by the model, in which it is the difference in skill in procedure C and in procedure N that affects the physician's choice. The rate of bad outcomes in each group proxies for surgical skill because the vast majority of high-risk women get C-sections and most very low-risk women do not. At the same time, because the very high-risk and very low-risk groups are defined only in terms of underlying medical risk factors, the measure is not contaminated by the endogeneity of the actual choice of C-section within these risk categories. This measure also exhibits considerable variation between doctors with a mean of -0.0493 (given that bad outcomes are more frequent in high risk cases than in low risk cases) and a standard deviation of 0.0646. The first percentile of this variable is -0.25, while the 99th percentile is 0.079. Again, we normalize this measure by calculating a Z-score for ease of interpretation.

The effects of decision making skill (from the estimated β_{jt} in equation 10) and our measure of procedural skill are presented in Table 9. The top part of the table reports the results of skill upon C-section rates. The formal model supposes that the distribution of outcome variables x is the same for all doctors. We control for this by also doing the analysis at the market level. In that case we are identifying market level variation in diagnostic skill to control for patient self-selection to physicians. The TSLS results refers to these two-stage least squares estimates that control for selection of patients to physicians at the market level. Notice that an increase in decision making skills leads to higher C-sections for the high risk patients, while it reduces the rate for low risk patients. More importantly, the effect of decision making skill has a zero average effect. This is important because most of the public policy concern has been with the high C-section rates, and not upon the quality of decision making.

The effect of decision making quality of the physician is reported in the lower part of the table. Notice that performance increases for both the high-risk and the low-risk groups. In other words, an *increase* in C-section rates for the high risk patients results in better outcomes. This effect is different than procedural skill, which mainly affects the level of C-sections via the α_j term in physician quality. We can see this because an increase in procedural skill increases the C-section rate for both high and low risk patients. However, in the lower panel we see that outcomes improve for both risk categories.

Our earlier work, Currie and MacLeod (2008), found strong and consistent effects of tort reform upon outcomes, consistent with the hypothesis that a C-section is not risk free, and that physicians

respond to financial incentives. These results are consistent with a long literature in health economics illustrating the relationship between financial incentives and procedure choice (e.g. Gruber and Owings (1996)). However, for the better physicians, the effect of these reforms were close to zero, consistent with our hypothesis that there are variations in physician quality, and that the better physicians are not affected by tort law (nor should they be - in the U.S. medical liability is a negligence regime, and hence only negligent physicians should respond to changes in the law).

More importantly, these results illustrate the role that diagnosis plays in determining patient outcomes, and that there is not a one size fits all approach for determining C-sections. We find that for low risk mothers the C-section rate is too high relative to the medically optimal level, while for high risk mothers it is *too low.* Currie and MacLeod (2013) conclude by observing:

> Taking the model to data on C-sections, the most common surgical procedure performed in the U.S., we show that improving diagnostic skills from the 25th to the 75th percentile of the observed distribution would reduce C-section rates by 11.7% among the low risk, and increase them by 3.8% among the high risk. Since in our application there are many more low risk women than high risk women, improving diagnosis would reduce overall C-section rates without depriving high risk women of necessary care. Moreover, we show that an increase in diagnostic skill would improve health outcomes for both high risk and low risk women, while improvements in surgical skill have much larger effects on high risk women. These results are consistent with the hypothesis that improving diagnosis through methods such as checklists, computer assisted diagnosis, and collaborative decision making could reduce unnecessary procedure use and improve health outcomes.

6 Conclusions

The paper outlines a human capital approach to measuring the treatment effect of choice in situations where it is not possible or practical to carry out trials of sufficient precision. I begin with a discussion of randomized control trials of drugs for treating depression. This example illustrates difficulty on measuring a consistent relationship between patient characteristics and outcomes. Thus, it is not surprising that Frank and McGuire (2000) find that the problems with health delivery for physical illness are all magnified when it comes to mental health. This is also consistent with the recent results of Dickstein (2012), who finds that prescription behavior by psychiatrists is very sensitive to the reimbursement rates offered by insurance plans. This points to a need for a better understanding of how to design treatment as a function of patient observables.

The rest of the essay discusses a human capital approach to this problem. It is built upon two generic features of human capital. First, the fact that experts have a great deal of training/human capital implies that their decisions can be used to to organize individuals into groups that as a

group should have similar treatment needs. Here simple machine learning techniques can used to estimate a propensity score for each group - the likelihood that individals in a group receive an intensive treatment by the average expert.

Second, even though experts are skilled, they necessarily make mistakes. This is consistent with the fact that human capital is expensive to acquire - at some point it is not worthwhile or possible to increase decision making skill. As emphasized by the Rubin/Holland potential outcomes approach, such errors are essential if we are to measure the size of a treatment effect. Under the hypothesis that conditional upon the propensity score, errors are uncorrelated with patient characteristics, then one can consistently estimate the treatment effect. This provides a "structural" interpretation propensity score estmators (Rosenbaum and Rubin (1983) and Hirano et al. (2003)).

The analysis also illustrates the point that when optimal decisions vary with the characteristics of the units, then the average treatment effect is not necessarily very useful (even if well measured) because it averages over a group of units where the treatment effect is both positive and negative. In the case of heart attack patients, Currie et al. (2016) find that the optimal choice from a medical point of view is to provide all patients with an invasive procedure. However, our results identify some systematic heterogeneity in treatment across patients. Physicians from better hospitals tend to be more responsive – namely, they are less likely to do an invasive procedure for low appropriateness patients, which in practice corresponds to older patients (see Table A1 in Currie et al. (2016) (could also say see Table 5)). This is consistent with the hypothesis that these physicians are sensitive to other factors than simply medical necessity when making their decisions.

In the case of child birth, Currie et al. (2016) find that there is a great deal of heterogeneity in the decision to perform a C-section. It is widely believed that some of this heterogeneity is due to financial incentives that may explain the high C-section rates in the United States.[17] We found this to be the case for low risk births. However, in the case of high risk births our results impy that the C-section rate is too *low.* When we average over the two groups, and take into account the number of women at risk, we find that the mean C-section rate in New Jersey is too low relative to the medically optimal rate.

Much more work is needed to explore the robustness of these results. However, the case of C-sections does illustrate an important public policy issue where more work is needed to link measured treatment effects to policy recommendation, a point that Heckman and Smith (1995) and Dehejia (2005) have already emphasized in the case of program evaluation. The finding in Currie et al. (2016) that average C-section rates are too low in New Jersey is consistent with recent work by Molina et al. (2015) who look at C-section rates world wide. They find that the WHO guidelines of 10%-15% C-section rates are too low, and that 19% may be a more appropriate norm. However, as D'Alton and Hehir (2015) point out in their discussion of this paper, whether or not to have a C-section should be based upon high quality information. Not only should the C-section

[17] See Gruber and Owings (1996) and Consumer Reports (2015).

incidence vary with the characteristics of the mother, it should also vary with the characteristics of the physicians and characteristics of the hospital where child delivery is occurring.

These examples provide concrete illustrations of what Deaton (2010) calls the well known "heterogeneity problem." The contribution of the human capital approach is to provide one way to combine structure with randomization, as recommended by Heckman (2010). Decision making by the expert provides structure to organizing patients into groups in a way that is analogous to the propensity score method of Rosenbaum and Rubin (1983). Once we condition upon best practice as perceived by the expert, then one can identify the conditional treatment effect under the hypothesis that even experts make mistakes. We can exploit the variation in error rates between experts to learn what strategies works best.

It is worth emphasizing that the approach here is a bit different from the typical machine learning strategy. For example, supervised learning of an algorithm begins with a training set produced by experts to "teach" the algorithm about what are the best decisions in certain situations. Once trained, one can test the algorithm out of sample (see Athey and Imbens (2015) for an explicit application of these ideas to estimating the conditional average treatment effect).

The approach suggested here combines the wisdom of experts to characterize sub-populations with the fact that experts do make mistakes (Kahneman and Klein (2009)). Rather than sample only the best decision makers, the human capital approach suggests using a large sample with many decision makers to generate variation in decisions over sub-populations of the treatment unit. This allows us to estimate the conditional average treatment effect for finer sub-populations than would be possible with structured randomized control trials. Within the medical community there has been a great deal of attention paid to improving decisions and reducing errors. One of the recognized challenges is to systematically collect more high quality data that would allow the type of the analysis suggested here.[18]

References

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996, Jun). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Angrist, J. D. and A. B. Krueger (1999). Empirical strategies in labor economics. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, pp. 1278–1357. Elsevier Science B.V.

Angrist, J. D. and J.-S. Pischke (2010, Spring). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

[18]See for example the discussion in Leape and Berwick (2005). See also Cipriani et al. (2016), who cite the lack of high quality, comprehensive data as hurdle to determine effective treatment for mental illness.

Athey, S. and G. Imbens (2015). Machine learning methods for estimating heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*.

Azorin, J., P. Llorca, N. Despiegel, and P. Verpillat (2003). Escitalopram is more effective than citalopram for the treatment of severe major depressive disorder. *L'Encephale 30*(2), 158–166.

Banerjee, A. and E. Duflo (2009). The experimental approach to development economics. *Annual Reviews 1*, 151–78.

Bech, P., P. Tanghøj, P. Cialdella, H. F. Andersen, and A. G. Pedersen (2004). Escitalopram dose–response revisited: an alternative psychometric approach to evaluate clinical effects of escitalopram compared to citalopram and placebo in patients with major depression. *International Journal of Neuropsychopharmacology 7*(3), 283–290.

Bose, S. S. and P. C. Mahalanobis (1938). On estimating individual yields in the case of mixed-up yields of two or more plots in field experiment. *Sankhyā: The Indian Journal of Statistics (1933-1960) 4*(1), 103–111.

Burke, W. J., I. Gergel, and A. Bose (2002). Fixed-dose trial of the single isomer ssri escitalopram in depressed outpatients. *Journal of Clinical Psychiatry*.

Chandra, A., D. Cutler, and Z. Song (2012). Chapter six - who ordered that? the economics of treatment choices in medical care. In T. G. M. Mark V. Pauly and P. P. Barros (Eds.), *Handbook of Health Economics*, Volume 2 of *Handbook of Health Economics*, pp. 397–432. Elsevier.

Chandra, A. and D. O. Staiger (2007). Productivity spillovers in health care: Evidence from the treatment of heart attacks. *Journal of Political Economy 115*(1), pp.103–140.

Charness, G. and P. Kuhn (2011). Lab labor: What can labor economists learn from the lab? In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics, Volume 4*, Volume 4. Elsevier.

Cipriani, A., X. Zhou, C. Del Giovane, S. E. Hetrick, B. Qin, C. Whittington, D. Coghill, Y. Zhang, P. Hazell, S. Leucht, P. Cuijpers, J. Pu, D. Cohen, A. V. Ravindran, Y. Liu, K. D. Michael, L. Yang, L. Liu, and P. Xie (June 8, 2016). Comparative efficacy and tolerability of antidepressants for major depressive disorder in children and adolescents: a network meta-analysis. *The Lancet*.

Consumer Reports (2015, February). Risks of c-sections.

Currie, J. and W. B. MacLeod (2008, May). First do no harm? tort reform and birth outcomes. *Quarterly Journal of Economics 123*(2), 795–830.

Currie, J., W. B. MacLeod, and J. V. Parys (2016, May). Physician practice style and patient health outcomes: The case of heart attacks. *Journal of Health Economics 47*, 64–80.

Currie, J. M. and W. B. MacLeod (2013, April). Diagnosis and unnecessary procedure use: Evidence from c-section. (18977). Forthcoming in Journal of Labor Economics.

Currie, J. M. and W. B. MacLeod (2017, January). Diagnosis and unnecessary procedure use: Evidence from c-section. *Journal of Labor Economics*. Forthcoming.

Cusin, C., H. Yang, A. Yeung, and M. Fava (2010). Rating scales for depression. In L. Baer and M. Blais (Eds.), *Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health*, Chapter 2, pp. 7–35. Springer.

D'Alton, M. E. and M. P. Hehir (2015). Cesarean delivery rates: Revisiting a 3-decades-old dogma. *JAMA 314*(21), 2238–2240.

Deaton, A. (2010, June). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics 125*(1–2), 141 – 173. Experimental and non-experimental evaluation of economic policy and models.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. New York, NY: Springer-Verlag.

Dickstein, M. J. (2012, April). Physician vs. patient incentives in prescription drug choice. Stanford University.

Donnelly, C. L., K. D. Wagner, M. Rynn, P. Ambrosini, P. Landau, R. Yang, and C. J. Wohlberg (2006). Sertraline in children and adolescents with major depressive disorder. *Journal of the American Academy of Child & Adolescent Psychiatry 45*(10), 1162–1170.

Emslie, G. J., D. Ventura, A. Korotzer, and S. Tourkodimitris (2009). Escitalopram in the treatment of adolescent depression: a randomized placebo-controlled multisite trial. *Journal of the American Academy of Child & Adolescent Psychiatry 48*(7), 721–729.

Epstein, A. J. and S. Nicholson (2009). The formation and evolution of physician treatment syltes: An application to cesarean sections. *Journal of Health Economics 28*, 1126–1140.

Fabre, L. F., F. Abuzzahab, M. Amin, J. Claghorn, J. Mendels, W. M. Petrie, S. Dube, and J. G. Small (1995). Sertraline safety and efficacy in major depression: a double-blind fixed-dose comparison with placebo. *Biological psychiatry 38*(9), 592–602.

Findling, R. L., A. Robb, and A. Bose (2013). Escitalopram in the treatment of adolescent depression: A randomized, double-blind, placebo-controlled extension trial. *Journal of child and adolescent psychopharmacology 23*(7), 468–480.

Fisher, R. A. (1936). Design of experiments. *British Medical Journal 1*(3923), 554.

Frank, R. G. and T. G. McGuire (2000). Economics and mental health. In M. V. Pauly, T. G. McGuire, and P. P. Barros (Eds.), *Handbook of Health Economics*, Volume 1, Part B, Chapter 16, pp. 893 – 954. Elsevier.

Freedman, D. A. (2006, DEC). Statistical models for causation - what inferential leverage do they provide? *Evaluation Review 30*(6), 691–713.

Gruber, J. and M. Owings (1996). Physician financial incentives and cesarean section delivery. *The RAND Journal of Economics 27*(1), pp.99–123.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. New York, NY: Springer.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature 48*(2), 356–98.

Heckman, J. J. and B. E. Honore (1990). The empirical content of the roy model. *Econometrica 58*(5), pp.1121–1149.

Heckman, J. J. and J. A. Smith (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives 9*(2), 85–110.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Imbens, G. W. and D. B. Rubin (2011). *Causal Inference in Statistics and Social Sciences*. Oxford University Press.

Johnson, E. M. and M. M. Rehavi (2016). Physicians treating physicians: Information and incentives in childbirth. *American Economic Journal: Economic Policy 8*(1), 115–41.

Kahneman, D. and G. Klein (2009). Conditions for intuitive expertise a failure to disagree. *American Psychologist 64*(6), 515–526.

Leape, L. and D. Berwick (2005). Five years after to err is human: What have we learned? *JAMA 293*(19), 2384–2390.

Lepola, U. M., H. Loft, and E. H. Reines (2003). Escitalopram (10–20 mg/day) is effective and well tolerated in a placebo-controlled study in depression in primary care. *International clinical psychopharmacology 18*(4), 211–217.

List, J. A. and I. Rasul (2011). Field experiments in labor economics. *Handbook of Labor Economics 4*, 103–228.

Llorca, P.-M., J.-M. Azorin, N. Despiegel, and P. Verpillat (2005). Efficacy of escitalopram in patients with severe depression: a pooled analysis. *International journal of clinical practice 59*(3), 268–275.

Ludwig, J., D. E. Marcotte, and K. Norberg (2009). Anti-depressants and suicide. *Journal of Health Economics 28*(3), 659–676.

Macleod, W. B. (2016). Human capital: The missing link between behavior and economics. *Labor Economics*. Address to Society of Labor Economists, June 2015, Montreal, Canada.

Mahalanobis, P. C. (1944). On large-scale sample surveys. *Phil Trans Roy Soc London Ser B Biol Sci 231*((584)), 329–451.

March, J. S., J. Biederman, R. Wolkow, A. Safferman, J. Mardekian, E. H. Cook, N. R. Cutler, R. Dominguez, J. Ferguson, B. Muller, et al. (1998). Sertraline in children and adolescents with obsessive-compulsive disorder: a multicenter randomized controlled trial. *Jama 280*(20), 1752–1756.

Molina, G., T. G. Weiser, S. R. Lipsitz, M. M. Esquivel, T. Uribe-Leitz, T. Azad, N. Shah, K. Semrau, W. R. Berry, A. Gawande, and A. B. Haynes (2015). Relationship between cesarean delivery rate and maternal and neonatal mortality. *JAMA 314*(21), 2263–2270.

Ninan, P., D. Ventura, and J. Wang (2003). Escitalopram is effective and well tolerated in the treatment of severe depression. In *Poster presented at the Congress of the American Psychiatric Association, May*, pp. 17–22.

Olie, J., K. Gunn, and E. Katz (1997). A double-blind placebo-controlled multicentre study of sertraline in the acute and continuation treatment of major depression. *European psychiatry 12*(1), 34–41.

Pigott, T. A., A. Prakash, L. M. Arnold, S. T. Aaronson, C. H. Mallinckrodt, and M. M. Wohlreich (2007). Duloxetine versus escitalopram and placebo: an 8-month, double-blind trial in patients with major depressive disorder. *Current Medical Research and Opinion 23*(6), 1303–1318.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), pp.135–146.

Schneider, L. S., J. C. Nelson, C. M. Clary, P. Newhouse, K. R. R. Krishnan, T. Shiovitz, and K. Weihs (2003). An 8-week multicenter, parallel-group, double-blind, placebo-controlled study of sertraline in elderly outpatients with major depression. *American Journal of Psychiatry 160*(7), 1277–1285.

Smith, J. A. and P. E. Todd (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics 125*(1–2), 305 – 353. Experimental and non-experimental evaluation of economic policy and models.

Song, Y., J. Skinner, J. Bynum, J. Sutherland, J. E. Wennberg, and E. S. Fisher (2010). Regional variations in diagnostic practices. *New England Journal of Medicine 363*(1), 45–53.

Stahl, S. M. (2000). Placebo-controlled comparison of the selective serotonin reuptake inhibitors citalopram and sertraline. *Biological psychiatry 48*(9), 894–901.

Ventura, D., E. P. Armstrong, G. H. Skrepnek, and M. Haim Erder (2006). Escitalopram versus sertraline in the treatment of major depressive disorder: a randomized clinical trial. *Current Medical Research and Opinion 23*(2), 245–250.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Wade, A., O. M. Lemming, and K. B. Hedegaard (2002). Escitalopram 10 mg/day is effective and well tolerated in a placebo-controlled study in depression in primary care. *International clinical psychopharmacology 17*(3), 95–102.

Wagner, K. D., P. Ambrosini, M. Rynn, C. Wohlberg, R. Yang, M. S. Greenbaum, A. Childress, C. Donnelly, D. Deas, S. P. D. S. Group, et al. (2003). Efficacy of sertraline in the treatment of children and adolescents with major depressive disorder: two randomized controlled trials. *Jama 290*(8), 1033–1041.

Wagner, K. D., J. Jonas, R. L. Findling, D. Ventura, and K. Saikali (2006). A double-blind, randomized, placebo-controlled trial of escitalopram in the treatment of pediatric depression. *Journal of the American Academy of Child & Adolescent Psychiatry 45*(3), 280–288.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Jour Exp Agric 1*((2)), 129–142.

Table 5: Patient Characteristics (X)

Appropriateness for Surgery:	**All**	**Low**	**High**
Female	0.40	0.53	0.27
Age	69.91	80.69	59.65
White	0.79	0.83	0.76
Black	0.08	0.07	0.10
Hispanic	0.10	0.08	0.11
Medicaid	0.04	0.02	0.06
Medicare	0.66	0.88	0.38
Private Insurance	0.21	0.07	0.39
Self Pay or Other	0.09	0.03	0.17
Morbidity Index	0.45	-1.33	2.02
Subsequent AMI	0.05	0.12	0.003
#Diagnoses	8.20	8.98	7.16
Arrhythmia	0.26	0.32	0.20
Hypertension	0.43	0.33	0.56
Congestive Heart Failure	0.32	0.51	0.11
Peripheral Vascular Disease	0.05	0.05	0.04
Dementia	0.03	0.09	0.00
Cerebral Vascular Disease	0.07	0.14	0.01
COPD	0.16	0.20	0.09
Lupus	0.02	0.03	0.01
Ulcer	0.01	0.01	0.00
Liver Disease	0.02	0.03	0.00
Cancer	0.06	0.10	0.02
Diabetes	0.21	0.18	0.22
Kidney Disease	0.15	0.28	0.03
HIV	0.003	0.004	0.002
N	658,553	217,323	223,853

Source: Table 2, Currie et al. (2016).

Table 6: Fraction of Estimated Provider Coefficients that are Significantly Different than $\beta = 1$ and $\alpha = 0$.

	Beta<1	Beta=1	Beta>1	Total
Alpha<0	0.028	0.138	0.010	0.176
Alpha=0	0.069	0.527	0.0096	0.606
Alpha>0	0.041	0.177	0.0007	0.219
Total	0.138	0.842	0.020	

N= 658,553 patients.

Source: Table 5b, Currie et al. (2016).

Table 7: Outcomes and Practice Style Among Patients with High and Low Appropriateness

	(1)	(2)	(3)	(4)	(5)	(6)
Appropriateness for Invasive Procedure:	**High**	**High**	**High**	**Low**	**Low**	**Low**
Outcome:	**Hosp. Aquired Infection**	**Died in Hospital**	**Discharged to Home**	**Hosp. Aquired Infection**	**Died in Hospital**	**Discharged to Home**
Low Responsiveness	0.007***	0.009***	-0.025***	-0.010***	-0.011***	0.008*
(Beta<1)	(0.002)	(0.001)	(0.003)	(0.003)	(0.003)	(0.004)
Low Aggressiveness	0.010***	0.009***	-0.019***	0.014***	0.013***	-0.024***
(Alpha<0)	(0.002)	(0.001)	(0.003)	(0.003)	(0.002)	(0.003)
High Aggressiveness	-0.003*	-0.005***	0.013***	-0.011***	-0.019***	0.021***
(Alpha>0)	(0.001)	(0.001)	(0.002)	(0.003)	(0.002)	(0.003)
Hospital*Year FE	Y	Y	Y	Y	Y	Y
Patient Appropriateness Index	Y	Y	Y	Y	Y	Y
Patient Age Categories & Gender	Y	Y	Y	Y	Y	Y
Previous AMI	Y	Y	Y	Y	Y	Y
Patient Comorbidities	Y	Y	Y	Y	Y	Y
Physician Characteristics	Y	Y	Y	Y	Y	Y
N	223853	223853	223853	217323	217323	217323
R^2	0.05	0.06	0.29	0.08	0.08	0.12

Notes: Source is Table 6, Currie et al. (2016). Standard errors are clustered at the provider level and shown in parentheses. * indicates p<0.05, ** indicates p< 0.01,*** indicates p<0.001. Alphas and Betas vary with each 3 years of physician experience. "Low appropriateness" indicates patient is below the 34th percentile of our index of appropriateness for invasive procedures. "High appropriateness" indicates patient is above the 66th percentile.

Table 8: Estimation of $\eta(x)$.

	All Doctors			Good Doctors Only		
	Coeff.	S.E.	Marginal Effect	Coeff.	S.E.	Marginal Effect
Age<20	-0.337	0.013	-0.075	-0.428	0.029	-0.095
Age >=25&<30	0.262	0.008	0.058	0.311	0.018	0.069
Age >=30&<35	0.434	0.008	0.096	0.483	0.017	0.107
Age >=35	0.739	0.009	0.164	0.840	0.018	0.186
2nd Birth	-1.347	0.007	-0.298	-1.448	0.015	-0.321
3rd Birth	-1.645	0.009	-0.364	-1.787	0.019	-0.396
4th or Higher Birth	-2.140	0.012	-0.474	-2.317	0.027	-0.513
Previous C-section	3.660	0.008	0.810	3.885	0.018	0.860
Previous Large Infant	0.139	0.029	0.031	0.293	0.065	0.065
Previous Preterm	-0.293	0.025	-0.065	-0.311	0.061	-0.069
Multiple Birth	2.879	0.014	0.638	3.278	0.032	0.726
Breech	3.353	0.016	0.742	3.810	0.040	0.844
Placenta Previa	3.811	0.054	0.844	3.843	0.116	0.851
Abruptio Placenta	2.048	0.030	0.454	2.196	0.072	0.486
Cord Prolapse	1.761	0.047	0.390	1.668	0.100	0.369
Uterine Bleeding	0.026	0.035	0.006	0.259	0.099	0.057
Eclampsia	1.486	0.096	0.329	1.047	0.230	0.232
Chronic Hypertension	0.745	0.025	0.165	0.754	0.060	0.167
Pregnancy Hypertension	0.639	0.013	0.142	0.696	0.029	0.154
Chronic Lung Condition	0.064	0.014	0.014	0.110	0.032	0.024
Cardiac Condition	-0.121	0.020	-0.027	-0.175	0.042	-0.039
Diabetes	0.558	0.011	0.124	0.547	0.025	0.121
Anemia	0.131	0.018	0.029	0.203	0.043	0.045
Hemoglobinopathy	0.116	0.047	0.026	0.067	0.092	0.015
Herpes	0.461	0.024	0.102	0.558	0.049	0.124
Other STD	0.052	0.017	0.012	0.064	0.039	0.014
Hydramnios	0.616	0.018	0.136	0.645	0.042	0.143
Incompetent Cervix	0.043	0.035	0.010	-0.119	0.093	-0.026
Renal Disease	-0.024	0.031	-0.005	-0.057	0.067	-0.013
Rh Sensitivity	-0.045	0.040	-0.010	-0.082	0.109	-0.018
Other Risk Factor	0.276	0.006	0.061	0.210	0.013	0.047
Constant	-1.414	0.007	-0.313	-1.374	0.015	-0.304
# Observations	1169654			262174		
Pseudo R^2	0.32			0.322		

Notes: Source, Table 1, Currie and MacLeod (2017)The model also included indicators for missing age, parity, and risk factors. The correlation between rho estimated using the two different models is .99.

Table 9: Effect of Physician Decision Making and Surgical Skill on P(C-section) and Health Outcomes

C-section Risk:	**OLS**	**OLS**	**OLS**	**TSLS**	**TSLS**	**TSLS**
	All	**Low**	**High**	**All**	**Low**	**High**
Dep. Var: C-Section						
Decision Making	0.004	-0.011	0.018	0.000	-0.016	0.019
	(0.002)	(0.002)	(0.002)	(0.006)	(0.005)	(0.008)
Procedural Skill Difference	0.003	0.003	0.003	0.020	0.017	0.030
	(0.002)	(0.001)	(0.002)	(0.010)	(0.008)	(0.011)
R-sq/Chi-sq.	0.410	0.044	0.321	710797	15293	62526
Dep. Var: Any Bad Outcome						
Decision Making	-0.008	-0.007	-0.009	-0.013	-0.013	-0.013
	(0.002)	(0.001)	(0.002)	(0.006)	(0.007)	(0.006)
Procedural Skill Difference	-0.017	-0.008	-0.027	-0.058	-0.047	-0.072
	(0.002)	(0.002)	(0.002)	(0.006)	(0.007)	(0.006)
R-sq/Chi-sq.	0.020	0.016	0.023	6750	13635	1695
# Observations	968748	469170	499578	968748	469170	499578

Notes: Source - Table 4, Currie and MacLeod (2017). Standard errors clustered at the 3-digit zip code level. Regressions also include market price, estimated C-section risk, indicators for African-American, Hispanics, race missing, education (less than high school, high school, some college, missing), married, married missing, Medicaid, Medicaid missing, teen mom, 25-34, 35 plus, smoking, smoking missing, male child, parity 2, parity 3, parity 4 plus, parity missing, month and year of birth indicators, indicators for 3-digit zip code, and an indicator for whether the birth was on a week day. R-squared shown for OLS and Chi-squared shown for TSLS.