COLUMBIA | SIPA
Center for Development Economics and Policy

CDEP-CGEG WORKING PAPER SERIES

CDEP-CGEG WP No. 26

Viewpoint: The Human Capital Approach to Inference

W. Bentley MacLeod

March 2016

COLUMBIA | SIPA
Center on Global Economic Governance

Viewpoint: The Human Capital Approach to Inference[1]

W Bentley Macleod
Columbia University and NBER

March 22, 2016

[1]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 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.

Abstract

The purpose of this essay is to discuss the "human capital" approach to inference. Observed decisions by experts can be used to organize data on their decisions using simple machine learning techniques. The fact that the human capital of these experts is heterogeneous implies that errors in decision making are inevitable, which 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 some data from medical decision making in the context of treating depression, heart disease, and adverse childbirth events.

"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 to modern empirical economics. First, there is research using structural models that begins by assuming 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)). The second approach addresses the self-selection of individuals 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)).

There is general agreement that explicit randomization provides one of the cleanest way to obtain a measure of the effect of choice. It is also certainly the approach that is growing in importance and popularity, as we can see in Table 1. Notice as well that along with the rise in empirical research we see a corresponding fall in theory. One goal of the approach outlined here is to consider new ways to use theory to explore data. Also, while the rise in the use of experiments in economics is quite recent, the development of experimental techniques has a long history in economic development. Some of the earliest work tackled the problem of improving agricultural production in developed (Yates (1933); Bose and Mahalanobis (1938)) and developing countries (Bose and Mahalanobis (1938)). Such experiments take a long time, and it was understood early on that one could could not rely only upon experimental methods. For example, Mahalanobis (1944) provides a wonderful discussion of survey techniques he developed to study agriculture in India under challenging conditions to supplement results from field experiments.

Regardless of whether one uses an experiment or survey techniques, the goal is to measure the effect of treatment, based upon either explicit or quasi randomization. In contrast, the structural approach follows a more "Popperian" strategy of using theory to create a formal model to organize the data, and then proceeds by systematic rejections of the model that hopefully lead to subsequent improvements in the theory.[1] The purpose of this essay is to suggest a way to combine these approaches using what I call the "human capital approach to inference."

The approach is intended to help us learn from data sets that involve decisions by experts.

[1]See Popper (1963). Though it is worth emphasizing that Popper (1957) did not in fact believe that economics could be a science! Popper's curse is the claim that human affairs are so complex that one cannot distinguish between good models and just so stories. See Boltanski (2014).

In this paper I will explicitly discuss medical decision making in the treatment of depression and heart attacks, as well as assisting in child birth. Kahneman and Klein (2009) discuss the two contrasting view on human decision making - essentially human decision making can be viewed has a half full and half empty glass. Gary Klein works on psychology of expert decision, whose goal is to understand how people are able to achieve amazing performance in so many areas. Humans carry out complex surgery, drive cars and airplanes, play complex games – domains where we are just beginning to have computer software that is able to carry out the task. In contrast, behavioral economics, pioneered by Daniel Kahneman, focuses on where we mess up: the many situations where humans face apparently simple decisions, yet make woefully poor decisions. The purpose of Kahneman and Klein (2009) is to highlight that they are both right (which is why they can agree) – the same person can simultaneously make both brilliant and really poor decisions.

This observation is consistent with the economist's notion of human capital. These are the set of productive skills that take time to acquire. Individuals with high levels of human capital are capable of making high-quality domain-specific decisions. The fact that human capital is *expensive* implies that at the margin they can always do better, and hence there will be situations where even a highly skilled individual will make mistakes. We also know from labor economics that in any large population there is great variety in skill level, and hence also variation in error making.

The human capital approach to inference can be used in situations where we have a large number of persons to be treated by skilled decision makers. If we were to do a randomized control trial (RCT), then individuals would be randomly allocated to treatment and control, and then we would compare the outcomes. The problem is that in many cases, particularly in medical decision making, the optimal treatment varies with the characteristics of the patient. For example, some individuals face adverse reactions to drugs, and others have a natural immunity to disease, leading to heterogeneous responses to both treatment and placebo. The potential variation is substantial, which is why physicians spend years studying different possible conditions, and associating them with the appropriate treatment.

Let us now suppose that in addition to having a large set of patients, information on their characteristics, treatment, and outcomes, we also have them matched to physicians, with a large number of patients for each physician. We then proceed by using the fact that these physicians are experts, and hence on average their treatments are helpful. Assuming that there are only two treatment choices, A or B, we can use the decisions by the physicians to organize the data by the probability that a physician chooses A for patient i with characteristics x_i. This yields a propensity score $\eta(x_i)$. This is a straightforward machine learning exercise – given features x_i, what is the likelihood that choice A will be made.

One approach to machine learning would be to stop at this point. Namely, use the data to build a model of how expert physicians make choices. There is a huge literature studying this problem.[2] For example, we can view the recent work to produce self-driving cars as one in which the machine is learning to be as good as a human at such a task. However, we can

[2]For example the early work on machine learning was funded by the U.S. Postal Office in an attempt to mechanize reading addresses on letters, a task that humans could do quite well. See Mori et al. (1992).

Table 1: Percent Distributions of Methodology of Published Articles, 1963–2011

	Type of study				
Year	Theory	Theory with simulation	Empirical: borrowed data	Empirical: own data	Experiment
1963	50.7	1.5	39.1	8.7	0
1973	54.6	4.2	37.0	4.2	0
1983	57.6	4.0	35.2	2.4	0.8
1993	32.4	7.3	47.8	8.8	3.7
2003	28.9	11.1	38.5	17.8	3.7
2011	19.1	8.8	29.9	34.0	8.2

Source: Hamermesh (2013).

do a bit more. Once we have the propensity score, then we can proceed, as in Rosenbaum and Rubin (1983), to estimate the effect of choice conditional upon the propensity score. We differ from the standard propensity score approach in two ways. The first, is that we are concerned with the *conditional average treatment effect (CATE)* - the effect of treatment conditional upon characteristics x_i. As individual characteristics change, the optimal choice may change. The hope is that if we make a choice conditional upon the score x_i, this can result in better outcomes on average for individuals with this score.

Second, the goal of the propensity score estimator is to provide better control for observable characteristics, and the endogenous selection of individuals based upon their characteristics into treatment. In our case, since we have information on who treats, we can use the fact that human capital is limited, and hence physicians not only make errors, vary in the frequency with which mistakes are made. This allows use to measure the effect of treatment conditional upon patient characteristics, or CATE, and physician identity. We can ask which physicians get better performance, and what are the characteristics of their decisions that achieve better outcomes. The rest of the paper fills in the details of this discussion.

The agenda is as follows. The next section briefly reviews the Rubin/Holland. Section 3 discusses the use of randomized control trials to estimate average treatment effects. The purpose of getting a good estimate of the treatment effect is to make a better decision. However, in many cases the average treatment effect is misleading due to large variations of the effect on sub-populations. In many cases, particularly in medicine, the complexity of the environment is such that randomizing over all the sub-populations of interest is simply impossible. As Melinda Beck points out in a recent Wall Street Journal article, genetic variation in metabolic rates can lead to large variation in response to drugs.[3] In this section data from RCT testing of anti-depression drugs is discussed to illustrate how the trials often have little relevance for decision making in the field. In particular, this example provides another illustration of the point that the average treatment effect is not always the most important parameter of interest (see for example the discussions in Heckman (2010) and

[3]Is Your Medicine Right for Your Metabolism?, Wall Street Journal, March 14th, 2016.

Deaton (2010)). Section 4 introduces the "human capital" approach to inference, and provides conditions under which valid inference is possible. Section 5 discusses two applications of this idea to heart attack treatment and whether or not to deliver a child by a C-section. The final section has some concluding remarks.

2 The Rubin/Holland Model

In this section I review the well known Rubin-Holland model outlined by Holland (1986) and explicitly link it to optimal decision making.[4] The question is how to use evidence from an experiment or observational data to make better decisions. I will reiterate the basic point by Holland (1986) that doing so requires some modeling assumptions. In practice these assumptions are typically implicit, rather than explicit, which in turn can muddy the relationship between theory and evidence.

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, or some other well defined set of observable features. Individuals may also be firms or countries, though for the current discussion we can think of them as a collection of persons denoted by:

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

where x_i is the characteristic of individual i. Here I deviate slightly from Holland where the primitive is typically the set U. 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 a special case where each person is a unique point in X.

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. We are implicitly making the *stable unit treatment value assumption (STUVA)* - the decision for unit $j \neq i$ does not effect the potential outcomes for unit i. The *average treatment effect (ATE)* of choice 1 is given by:

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

This is typically the parameter estimated with a randomized control trial (see Imbens and Rubin (2011)). One procedure is as follows. We randomly select from U - the set of individuals that match the criteria in set X - $2n$ individuals, who are randomly assign to group

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

A - U_A and group B - U_B. 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 $d_i = 0$ otherwise. 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 averge treatment effect:

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

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:*

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

Proof. We follow Deaton (2010). First:

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

Next observe that:

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

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:

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

□

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 $\tau\left(X\right)$ 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.

Table 2: Sales of SSRI drugs and mood stabilizers in the US

Drug name	Lexapro (Forest Laboratories)		Zoloft (Pfizer)		Abilify (Otsuka Pharmaceutical)		Lamictal (GlaxoSmithKline)	
Drug type	SSRIs				Mood Stabilizers			
	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 first condition, $\tau(X) = E\{\hat{\tau}(Data, n)\}$, 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 $\tau(X)$ as a function of different assignment mechanisms. In many cases, as both Deaton (2010) and Heckman (2010) observe, we are also interested in the treatment effect for sub-populations of X.

For example, consider the problem of choosing a drug for the treatment of depression. First note that this is a very significant question. 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 drug.

We can view trials as have having three outcomes, $u_i \in \{V, 0, -B\}$, where $V > 0$ is to feel well, 0 is to be depressed, and $-B < 0$ 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 $u_i = V$. 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:

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

Second, one needs an instrument to measure the outcome of the trial, which is necessarily different than the long term outcome. 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 (CRDS-R) to produce a score before and after treatment, y_i and $y_i^{d_i}$.[5]

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 Table 3.[6] 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 Table 3. 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, Ludwig et al. (2009) observe, these results lack external validity because the target individuals must, for ethical reasons, be excluded from the studies.[7]

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.[8]

The American Psychiatric Association looked at the question of how treatment affects suicide rates. The results are shown in Table 4. As one can see for the different ages groups the success 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 label" warnings to this effect. Given that by age 25 suicide has already claimed individuals, the positive effect at that age may due in part to the selection effect of suicide!

[5] See Cusin et al. (2010)

[6] Studies looking 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).

[7] 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%.

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

Table 3:

Study	Citation	# treatment	# placebo	Age	Dosage (mg/d)	Duration	RSR (p-value)	RRR (p-value)
			A) Lexapro (Escitalopram)					
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)	
			B) Zoloft (Sertraline)					
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.

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

Age Range	Drug-Placebo Difference in Number of Cases of Suicidality per 1000 Patients Treated
<18	14 additional cases
18-24	5 additional cases
25-64	1 fewer case
≥65	6 fewer cases

Notes: Results are from RCTs on all antidepressants for patients with MDD, obsessive compulsive disorder (OCD), or other psychiatric disorders. —

At the end, what we would like from these trials is to identify the set of characteristics X^+ such that $\tau(x_i) > 0, \forall x_i \in X^+$. This in turn determines the optimal rule $d_i^*(x_i) = 1$ if and only if $x_i \in X^+$. One way to do this is to estimate the conditional average treatment effect (CATE):

$$\tau^*(x) = E\left\{u_i^1 - u_i^0 | x_i = x\right\}, \tag{1}$$

this in turn directly answers the question of whether or not an individual with characteristics x_i should be treated or not. One could use an RCT for this if one could run a trial for all values of $x_i \in X$. In general that is simply impossible.

3 When is CATE better than ATE?

The previous illustrates that the ATE is not necessarily the most meaningful or important measure, particularly when the treatment effect varies in sign across the units. Given that the fundamental goal of many trials is to improve the quality of decision making, the purpose of this section is to outline how the machine learning literature evaluates decision making (here I rely heavily upon Devroye et al. (1996)). We begin by letting μ be a measure on X describing the distribution of individual characteristics. For each $x \in X$ the conditional average treatment effect (CATE) is defined by (1). Here the expectation is assumed to be computed at the time the decision is made.

The conditioning here may not be perfect, and there may be some heterogeneity that is not captured by x_i. However, we proceed under the hypothesis that x_i is the best information available at the time a decision is made, and we will be precise in evaluating the quality of decision making with the CATE relative to a decision that is made with the ATE. At the time a decision is made the treatment effect is uncertain, with an *ex ante* probability given by:

$$\eta^*(x) = Pr\left[u_i^1 - u_i^0 \geq 0 | i \in U, x_i = x\right]. \tag{2}$$

As we saw with the example from trials involving anti-depression drugs, the outcome scores have no cardinal meaning, and hence we suppose that decisions are based only upon an ordinal ranking of treatments. Heckman (2010) calls this a voting rule, though within the literature on pattern recognition it is called the optimal Bayes estimator for the best choice:

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

This maximizes the possibility of the "right choice" as opposed to the choice weights, the outcomes using u^1 and u^0, given the characteristics x_i of individual i.

This perspective provides a natural way to evaluate the quality of choice. Assuming that we have an optimal decision in hand (or at least it exists conceptually), we can use is to evaluate the quality of any other decision. Let $d : X \to \{0, 1\}$ be any measurable decision function, then the Bayes' risk is defined by:

$$\begin{aligned} L(d) &= Pr\left[d(x) \neq d^*(x)\right], \\ &= \mu\left(x | d(x) \neq d^*(x)\right). \end{aligned}$$

Note that $L(d^*) = 0$. The Bayes risk provides a measure of how frequently a choice deviates from the optimal choice. We can define choice for any feasible information set. Suppose Π is a measurable partition of X - namely $\forall A, B \in \Pi$, if $A \neq B$ then $A \cap B = \emptyset$ and $X = \cup A_{A \in \Pi}$. The optimal Bayes rule relative to this information set is:

$$d^*(x|\Pi) = E\left\{d^*(x) | \Pi\right\}.$$

In many contexts, such as in skill acquisition, this is a natural and important concept. For example, when doing surgery, there will be a complex sequence of steps. A new surgeon learns not by experimenting with the output (patient survives or not), but by comparing her choices at each step with the choices of her teacher. As a student, the instructor would be at her side prompting the next decision.

Theorem 2. *Given any decision function $d : X \to \{0, 1\}$ measurable relative to the partition Π then the Bayes risk satisfies:*

$$L\left(d^*(x|\Pi)\right) \leq L(d).$$

This result follows from an appropriate modification of Theorem 2.2 in Devroye et al. (1996).

The ATE corresponds to the case where Π consists only of the set X. Suppose the ATE is given by:

$$\tau(X) = \int_{x \in X} \tau(x)\, d\mu.$$

The corresponding ATE decision rule derived from this (and the one used in the case of SSRIs) is:

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

This is not the only decision rule possible, and it is not necessarily the best rule based upon the Bayes risk. Specifically, let:

$$\begin{aligned} \eta^{ATE} &= E\left\{d^*(x)\right\}, \\ &= Pr\left[d^*(x) = 1\right]. \end{aligned}$$

The optimal decision rule in this case is $d^{ATE*} = d^*(x|X) = 1$ if $\eta\left(I^{ATE}\right) \geq 1/2$ and zero otherwise. From theorem 2 we know that:

$$L\left(d^{ATE*}\right) \leq L\left(d^{ATE}\right).$$

Since there is no a priori assumptions made regarding the distribution of $\tau(x)$, then we can have $\tau(X) < 0$ while $\eta^{ATE} > 1/2$. If fact one has immediately the following result:

Proposition 3. *The optimal Bayes rule based upon the ATE, d^{ATE*}, has a strictly lower Bayes risk than a decision based upon the ATE, d^{ATE}, if and only if $\tau(X)\left(\eta^{ATE} - 1/2\right) < 0$.*

This result shows that decision making based upon the ATE may not be optimal in some cases. For example, suppose that the treatment effect is positive for a X^+, and given by $\tau^+ > 0$. While for $X^- = X - X^+$, we have $\tau^- < 0$. If $\mu(X^+) > \mu(X^-)$, and one has to choose the same choice for the whole population, the Bayes optimal choice is $d^{ATE*} = 1$. However, if

$$\tau^+\mu\left(X^+\right) + \tau^-\mu\left(X^-\right) < 0,$$

then the choice from using an RCT to measure the ATE would recommend $d^{ATE} = 0$, in which case:

$$\mu\left(X^-\right) = L\left(d^{ATE*}\right) < L\left(d^{ATE}\right) = \mu\left(X^+\right).$$

The problem is that one cannot determined d^{ATE*} without information on how $d^*(x)$ varies with x. We now turn to this issue.

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 the CATE. 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 $x_i \in X$. In the spirit of the SUTVA, I assume that physicians treat "in a bubble."[9] - 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 in our model is represented 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 (2015), 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 *learn.* 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):

$$\tau_{ij}(x) = \tau(x) + \epsilon_{ij},$$

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 human capital. I am assuming that the treatment effect is on a log scale, so that τ and y_i take values on $(-\infty, \infty)$. 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 $j \in J$ 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 US, 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:

$$\begin{aligned} Data &= \left\{ \left\{ x_i, u_i^{d_{ij}} | i \in U_j \right\} | j \in J \right\}, \\ &= \left\{ Data_j | j \in J \right\}. \end{aligned} \tag{3}$$

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

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

and X_j^0, and then carry out the decision:

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

What one learns in medical school are patient conditions that delineate 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. Suppose that $\tau(x) \in (-\infty, \infty)$, namely for all characteristics $x_i \in X$, the treatment effect is finite then the set of conditions where $d_j(x) = 1$ is given by:

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

In other words with a noisy signal there is always a chance a physician might recommends $d_i = 1$ for any $x_i \in X$! Hence, the best we can do is identify the probability that $d_i = 1$ as function of x_i.

The solution provided by the human capital approach 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 treatment effect is τ, then the probability of getting treatment 1 is:

$$\begin{aligned} e(\tau) &= Pr[d_i = 1 | \tau, I^*], \\ &= \sum_{j \in J} \rho_j F\left(\frac{\tau}{\sigma_j}\right). \end{aligned} \tag{4}$$

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

$$e'(\tau) = \sum_{j \in J} \rho_j(\tau) f\left(\frac{\tau}{\sigma_j}\right) / \sigma_j > 0. \tag{5}$$

This implies a 1-to-1 relationship between the probability of treatment and the treatment effect τ. This term 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 4. *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 function, 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 $LATE = \bar{\tau}$. □

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 we 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:

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

A simply computation implies:

Proposition 5. *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 two features of human capital that we exploit are, first, that agents $j \in J$ are skilled, and hence the treatment effect should be monotonic with the propensity score. Second, 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

The traditional approach in propensity score estimation involves the creation of well defined groups, within which assignment to treatment and control are independent of individual characteristics. The idea here is to use the fact that the 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, it is impossible to estimate the treatment effect because we only observe the optimal choice, not the counter-factual. In this section we discuss two papers that apply these ideas to physician decision making.

The same framework is used in both papers. In both cases 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.[10] 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 and standard machine learning model.[11] We begin by estimating:

$$\hat{\eta}(x_i) = Pr[d_i = 1] = F(\Gamma x_i), \tag{6}$$

where $d_i = 1$ indicates an invasive procedure, F is the logit function, and Γ is a vector of parameter estimates so that:

$$\Gamma x_i = \sum_K \Gamma_k x_{ik},$$

where k indicates a patient characteristic. We can 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 do this by defining an index of patient condition $s(x)$ by:

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

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

$$\hat{\eta}_j(x_i) = Pr[d_i = 1] = F(\alpha_{jt} + \beta_{jt} s(x_i)), \tag{7}$$

[10] 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.

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

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

$$\hat{u}_j^H = \frac{1}{n_j^H} \sum_{i \in U_j \cap U^H} u_i, \tag{8}$$

$$\hat{u}_j^L = \frac{1}{n_j^L} \sum_{i \in U_j \cap U^L} u_i, \tag{9}$$

where $n_j^l = |U_j \cap U^l|$ is the number of patients served by physician j in population $U^l, i \in \{L, H\}$. We can then ask if 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 6 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 (7) 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 not a consensus on how to treat these patients, at least for this sample. 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. 7 presents the results from

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

how variation in practice affects various outcomes for high and low appropriateness patients (versions of equations (8) and (9)). What is interesting is that more aggressive physicians get better outcomes. Also, low responsiveness physicians get worse outcome 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, 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 other factors other than the treatment effect of an intervention. This illustrates one of the benefits of the human capital approach to measuring the treatment

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

effect because it allows us to learn about other factors that influence a decision, that might not be apparent if one were to simply do a clinical trial.

5.2 Caesarean Sections

There is a great deal of concern that the C-sections rates in the United States is too high. In order to help mothers make better decisions, Consumer 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. While it is mechanically true that choosing a hospital with a low C-section rate results in a lower 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 this 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 regarding the quality of decision making in real time during the labor and delivery process.

Recently, Johnson and Rehavi (2016) find evidence that when the mother is a physician, then she has a lower C-section rate and gets a better outcome. Given that such a mother has more medical skill, then we should expect that she chooses physicians more carefully, and that the physician attending to her will be conscious of the implicit monitoring. In this case we have both a selection and incentive effect. The next question is whether or not it is possible to evaluate the quality of physicians?

Currie and MacLeod (2013) do this using the 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 6. 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

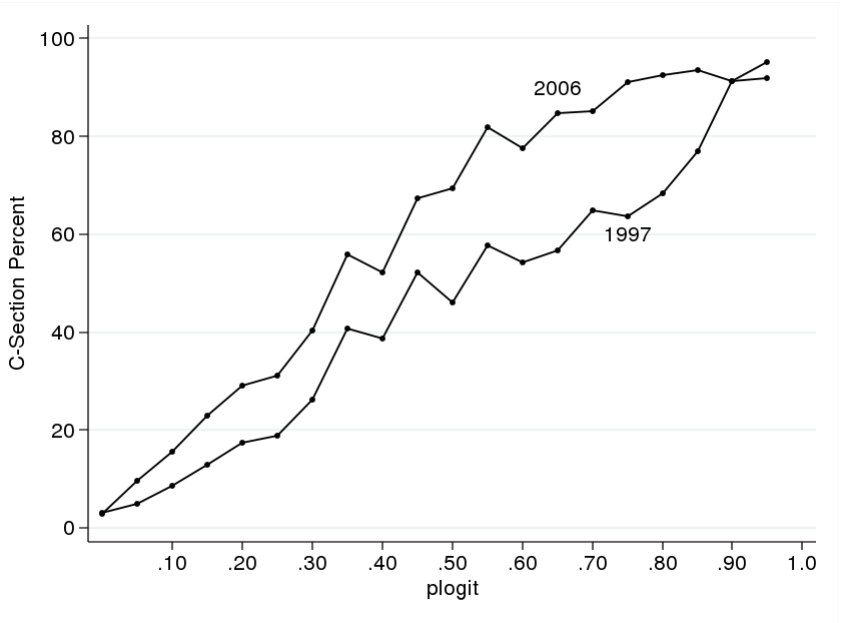
Table 7: Outcomes and Practice Style Among Patients with High and Low Appropriateness

	(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: 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.

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, though the largest increase occurs 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



The next question is whether physicians vary systematically in the way they treat patients. In Currie and MacLeod (2013) we provide a formal model of physician decisions that provides a structural interpretation of equation 7. 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 taking 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 7) 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. TSLS refers to our 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 inference with the goal of measuring the treatment effect of choice in situations where it is not possible or practical to carry out trials of sufficient precision. The example I use here is the problem of medical decision making. The underlying variation in human populations means that it is rarely the case that the choice of treatment, whether it is a drug or surgical procedure, will have a homogenous effects across the population at risk. For example, the same drug can be life saving for one person, while lethal for another.

The human capital approach begins with the hypothesis that we can use the decisions of experts to organize individuals into treatment groups that have similar characteristics, and hence the treatment effect within these groups is more homogeneous. Here machine learning techniques can be very useful because of their potential to categorize large amounts of data efficiently.

Second, even though experts are skilled, they necessarily make mistakes. Without mistakes there can be no learning - a randomized control trial is an extreme case of learning by forced randomization over possible treatments. In the context of physician decision making we can measure the variation in decision making, from which we can assess which physicians are getting better outcomes, which in turn allows us to evaluate the effect of treatment.

Third, the analysis illustrates a situation where the average treatment effect is not necessarily useful because it averages over a group of units where the treatment effect is both positive and negative. The machine learning approach focuses upon the quality of the decision making rather than the average value of a decision. In the case of a binary choice, when the expected value of the optimal decision is between 0 and 1, then we know we are in a situation with heterogeneous treatment effects.

This observation can help us interpret our findings. 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 among patients. In particular, 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). This is consistent with the hypothesis that 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 lead to excessively high C-section rates in the United States.[12] We found this to be the case for low risk births. However, in the case of high risk births 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. This provides another example of the problem with looking at the average treatment effect, which can mask significant heterogeneity in the optimal treatment choice.

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 to be 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 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 condition treatment effect under the hypothesis that even experts make mistakes. We can exploit the variation in errors rates between experts to learn what strategies works best.

Finally, I point out the well known fact that randomized control trials of drugs for treating depression are very inadequate. As Frank and McGuire (2000) point out, the problems with health delivery physical illness are all magnified when it comes to mental health. Given that there is little understanding on when a particular drug will work for a patient, treatment often involves psychiatrists doing their own mini-trials with each patient, and then adjusting medication as a function of the outcome. The starting point for treatment is typically the Diagnostic and Statistical Manual of Mental Disorders. What it does is attempt parse information about each patient, x_i, into categories, such "Schizophrenia Spectrum and Other Psychotic Disorders" or "Bipolar and Related Disorders". For each category there will be recommended treatments that are used as starting points. The DSM is careful to point out that these categories are not always definitive and that finding the right treatment can be difficult. At the moment it is simply not possible to apply something like the human capital approach to mental illness due to a lack of data. However, if there were large scale systematic

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

collection of data on individuals treated with mental illness, then progress could be made.

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. Of course, much work remains to refine the approach, and explore the extent to which "learning from our mistakes" can be automated to help improve decision making.

References

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed. ed.). Washington, DC: Autor.

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 C. D. (Eds.), *Handbook of Labor Economics*, pp. 1278–1357. Elsevier Science B.V.

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.

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.

Boltanski, L. (2014). *Mysteries and Conspiracies: Detective Stories, Spy Novels and the Making of Modern Societies*. John Wiley & Sons.

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*.

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.

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. Technical Report 18977, NBER, Cambridge, MA. Forthcoming in Journal of Labor Economics.

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.

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.

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.

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how?. *Journal of Economic Literature 51*(1), 162 – 172.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Larning*. 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.

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.

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. (2015, June). Human capital: The missing link between behavior and economics. Address to Society of Labor Economists, 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.

Mori, S., C. Y. Suen, and K. Yamamoto (1992). Historical review of ocr research and development. *Proceedings of the IEEE 80*(7), 1029–1058.

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.

Popper, K. R. (1963). *Conjectures and Refutations: the Growth of Scientific Knowledge*. New York, NY.: Basic Books.

Popper, K. R. (2002 (first published 1957), September). *The poverty of historicism*. Routledge.

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.

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 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 R2	0.32			0.322		

Notes: 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 Doctor Physician Making and Surgical Skill on P(C-section) and Health Outcomes

C-section Risk:	OLS All	OLS Low	OLS High	TSLS All	TSLS Low	TSLS High
Dep. Var: C-Section						
Decision Making	0.004	-0.011	0.019	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	0.319	230000	12674	88123
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	6600	13213	1721
# Observations	968748	469170	499578	968748	469170	499578

Notes: 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.