# Columbia University

*Department of Economics*
*Discussion Paper Series*



# Diagnosing Expertise: A Two Factor Model of Physician Skill



*Janet Currie and W. Bentley MacLeod*





*Discussion Paper No.: 1314-21*





***Department of Economics***
*Columbia University*
*New York, NY 10027*





November 2014

# Diagnosing Expertise: A Two Factor Model of Physician Skill*

Janet Currie
Princeton University and NBER
W. Bentley MacLeod
Columbia University and NBER

December 1, 2012
Revised, October 21, 2014

**Abstract**

This paper develops and applies a model in which doctors have two dimensions of skill: diagnostic skill and skill performing procedures. Higher procedural skill increases the use of intensive procedures across the board, while better diagnostic skill results in fewer intensive procedures for the low risk, but more for the high risk. Deriving empirical analogues to our theoretical measures for the case of C-section, we show that poor diagnosticians can be identified in the data and that improving diagnostic skill would reduce C-section rates by 15.5% in the bottom half of the risk distribution, and increase them by 5.5% in the top half. Such an change in the allocation of procedures would improve birth outcomes among all women.

*We thank Samantha Heep, Dawn Koffman, Jessica Van Parys and Geng Tong for excellent research assistance, and Amitabh Chandra, Jonathan Gruber, Amy Finkelstein, Kate Ho, Robin Lee, Jonathan Skinner and seminar participants at Princeton, Georgetown University, Havard Medical School, Kyoto University, NYU, the Japanese National Institute of Population and Social Security Research, Warwick University, University College London, the London School of Economics, the Paris School of Economics, the NBER Summer Institute, and the University of Michigan for helpful comments. This research was supported by a grant from the Program on U.S. Health Policy of the Center for Health and Wellbeing.

# 1 Introduction

A long literature focuses on experts' decision making and highlights the fact that "professional intuition is sometimes marvelous and sometimes flawed" (Kahneman and Klein, 2009). At the time a decision is rendered, its quality is difficult to evaluate, yet a flawed decision can have enormous consequences. The purpose of this paper is to measure the quality of decision making in the context of a well defined setting - the decision of an obstetrician to deliver a baby by either cesarean section or natural delivery. We show that this choice can be modeled as an information processing problem, and show further that physicians vary systematically in their decision making ability. While much of the literature in health economics deals with flawed decision making resulting in medical errors, over or under use of procedures, and poor health outcomes (Garber and Skinner 2008; Chandra et al., 2012), remarkably little previous research has considered expert decision making, or diagnosis, as a distinct aspect of medical practice.

This paper adds diagnostic skill to the canonical model of utility-maximizing physician behavior. Better diagnostic skill is conceived of as an ability to use patient information to improve the match between patient condition and the treatment they receive. This perspective highlights the ways that physicians adjust their behavior in response to the condition of the patient. We apply the model to a common binary decision faced by physicians: Whether to deliver a baby by C-section or not. In this context, physicians who are better at diagnosis relative to their peers should have *higher* C-sections rates for high risk patients and *lower* C-section rates for low risk patients.

The case of C-section is interesting for a number of reasons. First, there is widespread recognition that C-section rates vary across hospitals in ways that cannot be explained either by the condition of the patients or by their preferences (Kozhimannil et al. (2013)). The large rise in C-section rates over the past 15 years (from 1 in 5 to 1 in 3) has led to many proposals to lower them. For example, on January 1, 2014, the Joint Commission that provides hospital acreditation and allows hospitals to participate in the Medicaid and Medicare programs implemented a measure aimed at encouraging hospitals to reduce C-section rates among first time mothers with single, head-down fetuses. The Commission will publish a target rate based on a national sample of hospitals every quarter, and will require hospitals to publish and track their own rates in order to create pressure on them to lower rates. (Commission (2014) - see measure PC-02). However, our analysis suggests that public policies that focus on simply lowering C-sections rates could easily lead to worse outcomes if physicians who are good at diagnosis are encouraged to lower their C-section rates across the board.

In addition to the new focus in this paper on physician diagnostic skill, we also consider a more conventional type of physician skill, which we term procedural or surgical skill. We show that physicians who are good at surgery will have higher C-section rates for all patients, since their patients are at lower risk of bad outcomes. Over time C-sections have become safer, and hence the model implies that some increase in C-section rates over time is warranted and to be expected upon purely medical grounds.

Physicians may be more or less skilled at performing procedures, but they may also be more or less skilled at diagnosis. We show that these two dimensions of skill can be differentiated both theoretically and empirically, and that doctors whose decisions are more sensitive to information about a patient's medical conditions produce better health outcomes. In contrast to a model with one dimension of skill, our model implies that improvements in diagnostic skill make all patients better off, even as they reduce the unnecessary use of intensive procedures in patients for whom low intensity procedures are more appropriate. Moreover, we clearly show that there is variation in the way that physicians use the same information about patient condition, so that there can be no presumption that market forces alone (e.g. adjusting prices) will lead physicians to make the best decisions for their patients.

Applying our model to data on all deliveries in New Jersey from 1997 to 2006, we find that when diagnostic skill increases by one standard deviation, C-section rates fall 15.5% for women in the bottom half of the risk distribution, but rise 5.5% among women in the high risk half of the distribution. Given that there are many more C-sections among the high risk to begin with, such a change would both reduce the number of unnecessary surgeries performed on low-risk women, and increase the overall number of C-sections performed.

Thus, a surprising implication of our analysis is that not only are there too many C-sections being performed on low-risk women, but there are too few C-sections being performed on high-risk women. A one standard deviation improvement in diagnosis leads to reductions in the probability of a negative health outcome for all women: There is a reduction of 15.3% among the low risk, and of 9.1% among the high risk. When we further divide bad health outcomes into those that are bad for the mother and those that are bad for the infant, we find that reductions in bad outcomes among mothers are concentrated in the low risk (who become less likely to suffer the consequences of unnecessary surgeries), while for infants bad outcomes are reduced across the board. The one exception is neonatal death, which declines with better diagnosis only among the high risk (suggesting that C-sections are indeed life-saving among infants born to the highest risk mothers).

Contrasting the effects of diagnostic skill and surgical skill, we find that a one standard deviation improvement in surgical skill would increase the incidence of C-section 16.5% among patients in the lower half of the risk distribution and by 8.7% among patients in the upper half. The same change is estimated to reduce the incidence of any bad health outcome by 55.3% among the low risk, and by 50.4% among the high risk.

One might conclude that it is more important to improve surgical skill than diagnostic skill. But it may be considerably easier to improve diagnosis than to make bad surgeons into good ones. Indeed, policies such as checklists, computer aided diagnosis, or administrative structures that require that physicians seek approval before scheduling C-sections in women without risk factors, could perhaps be used as methods of improving diagnosis ((Baker et al., 2008); (Doi, 2007); (Gawande, 2009)). Our results suggest that with common procedures like C-section, it may well be possible to use existing administrative health data bases to identify doctors who are weak in terms of diagnosis and to make changes in the allocations of procedures that will improve patient health outcomes.

The rest of the paper is laid out as follows. Section II briefly reviews some of the relevant literature. The model is developed in Section III, which also explains how we use the observable administrative data to implement the model. Briefly, we first use the observable data to construct a measure of each patient's appropriateness for C-section. We then estimate doctor-specific regressions of the propensity to perform a C-section on this measure of appropriateness. This procedure yields an intercept and a slope term for each doctor, and the model explains the circumstances in which the estimated slope can be interpreted as a measure of the doctor's diagnostic skill. We also propose a proxy for the doctor's surgical skill. Section IV presents the results, and this is followed by a discussion and conclusions in Section V.

## 2 Background

Health care is an important area in which we all rely on experts, to make judgments that are referred to as diagnoses, and then to carry out the chosen procedure. Hence, it is not surprising that many studies of expertise have focused on physicians. Meehl (1954) reviewed a number of studies, mainly of clinical psychologists, and compared their forecasts to those generated by simple statistical models, including optimal linear combinations of variables that the experts also observed. He argued that predictions based on these simple models were generally more accurate than those of the experts. A more recent meta-analysis of 136 studies in clinical psychology and medicine also found that algorithms tended to either out-perform or to match the experts (Grove et al., 2000).

Kahneman and Klein (2009) argue that algorithms are most useful when we have confidence in the list of variables to be used for prediction; when we have a reliable and measurable outcome; when there is a large body of similar cases; when the cost/benefit ratio warrants the investment in developing an algorithm; and when the situation is sufficiently stable that the algorithm will not immediately become obsolete. The case of C-section appears to satisfy all of these criteria as we will argue further in the data section below. In the psychological studies discussed above, the experts and the statisticians generally had access to the same data. The advantage of the algorithms arises mainly because the algorithms are more consistent than the experts. However, in our application, we have another advantage which is that in our administrative birth records we observe the universe of cases over a given time period, whereas each doctor observes only their own cases.

Another difference between our study and many of those in psychology is that we are agnostic about the source of the "errors" in decision making. The psychology literature is concerned about whether the errors arise from factors such as over-confidence, or other heuristic biases. We are concerned with doctors who, for a variety of possible reasons, do not make the best use of the information at their disposal to make good decisions. The literature in health economics offers many possible reasons for these "mistakes." One common explanation for faulty decision making is "defensive medicine," the idea that doctors perform unnecessary procedures in order to protect themselves from lawsuits. However, Baicker et al. (2007) argue that there is little connection between malpractice liability costs and physician treatment of Medicare patients, while Dubay et al. (1999) and Currie and MacLeod (2008) cast doubt on the idea that physicians perform unnecessary C-sections primarily due to fear of lawsuits.

There is more evidence that physician decision making is swayed by financial incentives. The fee for performing C-sections exceeds the fee for performing natural deliveries. Gruber and Owings (1996) and Gruber et al. (1999) show that the incidence of C-section increases with the wedge between the two fees. Johnson and Rehai (2014) add to this literature by showing that financial incentives affect the treatment of non-physicians, but have no impact on the treatment of physician-patients, who are presumably better informed, and therefore less likely to meekly tolerate unnecessary procedures.

A third possibility is that doctors are influenced by the decisions of those around them. Chandra and Staiger (2007) study the choice of surgery vs. medical management of cardiac patients. Knowledge spillovers are the main theoretical driver of small area variation in procedure use in their model. Physicians in areas that specialize in surgery are assumed to become better at surgery and worse at medical management, and vice-versa. Their model raises the possibility of mismatch between patients and physicians. All patients in high surgery areas will be more likely to have surgery, even if medical management would be more appropriate for some of them.

Both Epstein and Nicholson (2009) and Dranove et al. (2011) investigate the prevalence of spillovers in the case of C-section and neither find much evidence for them: There is no convergence in practice styles among physicians in the same hospitals over time. And since C-section is often considered a rather simple surgery, the benefits from specialization may also be muted. Still, the model we discuss below is not inconsistent with the potential existence of either specialization or spillovers as practice presumably does help, and doctors could learn both to be better diagnosticians and better surgeons from observing their colleagues.

The most important insight from the Chandra and Staiger (2007) model may be that a reduction in the use of surgery in high use areas cannot be Pareto improving because patients who are good candidates for surgery will be harmed by a decline in the skill level of the physicians serving them. We will also argue that an across-the-board cut in C-section rates cannot be optimal because such a reduction will reduce the probability that high-need mothers will receive a procedure. What is desirable instead, is a reallocation of C-sections from low-need to high-need mothers.

Patient preferences are often cited as a fourth potential reason for medically unnecessary procedure use. In an innovative study using vignettes from patient and physician surveys, Cutler et al. (2013) assess the hypothesis that regional variations in procedure use are driven by differences in patient demand across areas. They conclude that patient demand is a relatively unimportant determinant of regional variations and that the main driver is physician beliefs about appropriate treatment that are often unsupported by clinical evidence. Similarly, previous studies have found little evidence that patient demand is driving the large differences in C-section rates across providers. (McCourt et al. (2007))

Finkelstein et al. (2014) address the same question using longitudinal Medicare claims data that allow them to track the same patients as they move through different health care markets. They suggest that about half of the observed variation in procedure use is due to supply-side factors, while half is due to patient-level, or demand-side factors. However, they conclude that much of the variation in patient demand is driven by exogenous patient health, and so probably does not simply reflect patient tastes for procedures. These findings agree with those of Cutler et al. (2013) in suggesting that patient preferences play a relatively small role in explaining variations in care.

These are all important lines of inquiry, and our theoretical model incorporates the effects of prices, patient preferences, patient health, and physician beliefs. However, our main focus is on identifying doctors who, for whatever reason, are making poor use of the observable data about their patients when making treatment decisions. We will show that patients of these doctors tend to have worse outcomes than other comparable patients. The fact that these doctors can be identified using simple models based on administrative data is relevant for policy because it suggests that it would be possible to improve patient outcomes by incorporating aids to diagnostic decision making into standard practice.

# 3 Framework

## 3.1 Overview

We begin by estimating a qualitative choice model using all of the data for the state of New Jersey between 1997 and 2006 following Smith et al. (2004). They show that a logistic model provides a clinically useful summary of factors related to C-section risk, and recommend providing this information in a useful fashion to the physician:

$$Prob\left\{C_i = 1\right\} = F\left(\beta X_i\right). \tag{1}$$

We use the model to measure standard treatment in New Jersey, and then explore how physicians deviate from this standard.[1]We then use the model to construct a one-dimensional measure of the patient's appropriateness for C-section:

$$h_i^I = \beta X_i. \tag{2}$$

By construction, our measure $h_i^I$ is positively correlated with patient conditions that are likely to merit a C-section. An important issue is that our constructed measure captures the standard of practice in New Jersey, but does not necessarily capture

[1]An alternative way to risk adjust would be to look at C-section rates for discrete cases. However, we have measures of 30 different conditions that all have some impact on the probability of C-section. With $n = 30$ there are $2^{30} = 1,073,741,824$ possible cases making it obviously impossible to risk adjust in this case-specific fashion!

appropriateness for a C-section in an absolute sense. Ideally, one might choose to construct $h_i^I$ using only "good doctors." As we will show below, there seems to be a good deal of consensus on the ranking of different patients by appropriateness for C-section in our data.

For each doctor $j \in J$ we estimate a model of the form:

$$Prob\{C_{ij} = 1\} = F\left(\theta_j h_i^I + \gamma_j)\right).$$

By including both a fixed effect $\gamma_j$ and a slope term $\theta_j$, this formulation provides a straightforward extension of the standard risk adjustment framework which treats differences between doctors as a doctor specific fixed effect (Epstein and Nicholson (2009), Srinivas et al. (2010)). We will show that there is variation in the slope term across the population of physicians, and that this variation is clinically significant in that it is associated with variation in outcomes. Before proceeding with the empirical analysis, we show that the slope and intercept terms can be viewed as providing a reduced form measure of information processing ability and surgical skill. In particular, the slope will be shown to vary only with variations in information, and not with surgical skill.

We let $h_i$ represent the true underlying condition of the patient and then suppose that our estimate $h_i^I$ (from equation 2) satisfies:

$$h_i^I = h_i + \epsilon_i^I, \tag{3}$$

where the error term has variance $\sigma_I^2$. The physician also has a signal of patient condition $h_i$, and the precision of this signal is what we use as a measure of *diagnostic skill.* We will show that this measure of diagnostic skill is positively related to the slope term $\theta_j$, whereas surgical skill affects the intercept term, $\gamma_j$, but not $\theta_j$.

### 3.2 Modeling Physician Decision Making

We begin with the standard hypothesis that physicians maximize their utility, but care about patient outcomes (Gaynor et al. (2004), Arlen and MacLeod (2005), Currie and MacLeod (2008) and Chandra et al. (2012)). As such the physician does her best to process the information she receives and then make the appropriate choice of procedure. The physician chooses between two procedures, $T \in \{N, C\}$ which generate the following physician payoffs:

$$
\begin{aligned}
u_{ij}(N) &= h_i^N + s_j^N + m_j^N \left(P^N\right) + \epsilon_{ijN}, \\
u_{ij}(C) &= h_i^C + s_j^C + m_j^C \left(P^C\right) + \alpha_j^P h_i^P + \epsilon_{ijC}.
\end{aligned}
$$

The term $h_i^T$ is an index of the health status of the patient $i$ in (log) utility terms after procedure $T$ is carried out, $s_j$ is the skill of the physician $j$ at performing procedure $T$, and $P^T$ is the cost of the procedure.[2]

The term $h_i^P$ represents a patient preference for procedure C (if it is negative, then she prefers procedure N).[3] The extent to which a physician is willing to alter her choice of procedure in response to the preferences of the patient during childbirth is denoted by $\alpha_j^P$, which represents *physician patient sensitivity,* the extent to which the physician responds to the preferences of the mother.[4] In what follows, we do not observe $h_i^P$, and this term can thus also be thought of as incorporating any other variables that are observed by the physician, but unrecorded in the data.

Given information $I_{ij}$ the physician chooses $C$ if and only if:

$$
E\left\{u_{ij}(C) - u_{ij}(N) | I_{ij}\right\} \geq 0. \tag{5}
$$

We suppose that surgical skill and prices are known and that our health measures have been normalized so that $E\left\{\epsilon_{ijC} - \epsilon_{ijN}\right\} = 0$. These assumptions allow us to restate the physician decision expressed in (5). The physician chooses the intensive procedure $(T = C)$ if and only if:

$$
E\left\{h_i | I_{ij}\right\} + s_j + m_{jt} + \alpha_j^P h_i^P \geq 0, \tag{6}
$$

[2]It is assumed that we have taken logs of level variables and hence utility is any real number (positive or negative), and the units have been defined appropriately.

[3]We could put these preference terms into both equations, but ultimately we are concerned about the relative preference of procedure C to N, and so we need only place this term into one equation.

[4]Note that this linear model can be generated from the a model that allows for complementarities:

$$
U_{ij}(T) = (H_i^T)(S_j^T) M_j^T \left(P^T\right), \tag{4}
$$

where $S_j^T$ is the skill of physician $j$ at doing procedure T and $M_j\left(P^T\right)$ is the expected pecuniary consequence of this choice as a function of the price paid, $P^T$ for procedure $T$. Taking logs yields:

$$
\begin{aligned}
u_{ij}(T) &= log(U_{ij}(T)) \\
&= log\left(H_i^T\right) + log\left(S_j^T\right) + log\left(M_j^T\left(P^T\right)\right) \\
&= h_i^t + s_j^T + m_j^T\left(P^T\right).
\end{aligned}
$$

where $s_j = s_j^C - s_j^N$, $m_j = m_j\left(P^C\right) - m_j\left(P^N\right)$ , and $h_i = h_i^C - h_i^N$. For simplicity, we normalize $h_i^C = 0$, so that $h_i = -h_i^N$. The term for technical skill $(s_j)$ increases with skill at $C$ , and decreases with skill at $N$. The term $m_j$ represents the relative cost of procedures C and N. Increases in the price of procedure $C$ is expected to increase $m_j$, while an increase in the price of procedure $N$ would decrease this term.

Suppose that the physician has prior beliefs regarding the patient's true condition $h_i$ which are given by $h_i \sim N\left(h_j^0, \sigma_j^2\right)$. If the mean satisfies $h_j^0 + s_j + m_j > 0$ , then the physician believes that most women in her practice should be getting a C-section.

Suppose that all physicians see the same basic information, $X_i$, and that variation in observed procedure choices reflects both their prior beliefs, $h_j^0$, and their ability to process information. The variance of prior beliefs, $\sigma_j^2$ , represents uncertainty about the appropriate choice. We can also define:

$$B_j = \frac{1}{\sigma_j^2}.$$

Formally this is the precision of the prior distribution and represents the extent to which a physician alters her beliefs in the face of new information. When $B_j$ is large ($\sigma_j^2$ is small), then the physician has strong prior beliefs that makes her less sensitive to the new information in $X_i$.

Given these beliefs, the physician observes the patient's condition and makes an assessment of her health status which is given by:

$$h_{ij} = h_i + \epsilon_{ij}, \tag{7}$$

where $\epsilon_{ji}$ is normally distributed with mean zero and variance $\sigma_{Dj}^2$. We model variability in the diagnostic skill of doctors using variation in $\sigma_{Dj}^2$. Formally, we define the *diagnostic skill* of a physician as:

$$D_j = \frac{1}{\sigma_{D_j}^2.}.$$

When $D_j$ is higher, the physician makes a more accurate estimate of the patient's condition $h_i$. Given these definitions we have:

**Proposition 1.** *Given a doctor's prior beliefs about the patient's condition $h_j^0$, the strength of the physician's prior beliefs, $B_j$, diagnostic skill $D_j$, and her information about the patient's condition, $h_{ij}$, then her* medical assessment *of a patient's condition* *is given by:*

$$E\left\{h_i|I_{ij}\right\} = \pi^0 h_j^0 + \pi^h h_{ij}$$

*where* $\pi^0 = \frac{B_j}{B_j+D_j}$ *and* $\pi^h = 1 - \pi^0 = \frac{D_j}{B_j+D_j}$.

The proof of this and subsequent propositions are in the appendix. This result follows directly from the optimal updating rule for normally distributed random variables (see DeGroot (1972)). Notice that physicians with higher diagnostic skill are more responsive to new information, and less dependent on prior beliefs.

The final piece of data used by the physician is the patient's preference for a C-section given by $h_i^P$. Suppose that patient preferences follow an arbitrary distribution $h_i^P \sim N\left(\bar{h}_j^P, \sigma_{Pj}^2\right)$, where $\bar{h}_P$ and $\sigma_{PJ}^2$ are practice specific parameters that can also affect the observed decision.

This decision model illustrates that there are at least five physician characteristics that affect decision making, which can be summarized by $\omega_{Dj} = \left\{s_j, h_j^0, B_j, D_j, \alpha_j^P\right\}$- physician surgical skill, prior beliefs about patient condition, the strength of these prior beliefs, diagnostic skill, and the parameter from the doctor's utility function describing how sensitive the physician is to patient preferences. Unobserved practice characteristics are given by $\omega_{Pj} = \left\{\bar{h}_j^P, \sigma_{Pj}^2\right\}$. Let $\omega_j = \{\omega_{Dj}, \omega_{Pj}\}$ denote the full set of physician and practice level characteristics.

If we substitute these expressions into equation 6 we can show that procedure $T = C$ is chosen by physician $j$ for patient $i$ if and only if:

$$T\left(h_{ij}, h_i^P|\omega_j\right) = \pi^0 h_j^0 + \pi^h h_{ij} + s_j + m_j + \alpha_j^P h_i^P \geq 0. \tag{8}$$

We can now derive the probability that a patient will receive procedure C as a function of her underlying condition $h_i$. Procedure C is chosen iff:

$$h_i + \frac{\pi^0 h_j^0 + s_j + m_j + \alpha^P \bar{h}_j^P}{\pi^h} \geq -\left(\epsilon_{ij} + \alpha_j^P \epsilon_j^P / \pi^h\right), \tag{9}$$

where $\epsilon_j^P$ is defined as the variation from the mean of patient preferences - $\left(h_j^P - \bar{h}_j^P\right)$. We can rewrite the second term of this equation as:

$$\begin{aligned} \gamma_j &= \frac{\pi^0 h_j^0 + s_j + m_j + \alpha_j^P \bar{h}_j^P}{\pi^h}, \\ &= \frac{B_j}{D_j}\left(h_j^0 + \bar{\gamma}_j\right) + \bar{\gamma}_j, \end{aligned} \tag{10}$$

where $\bar{\gamma}_j = s_j + m_j + \alpha_j^P \bar{h}_j^P$ are physician specific characteristics that are not part of physician expectations. Let us define:

$$\zeta_{ij} = -\left(\epsilon_{ij} + \alpha_j^P \epsilon_j^P / \pi^h\right),$$

which is a normally distributed random variable with mean zero and variance:

$$\sigma_{j\zeta}^2 = \left(\sigma_{Dj}^2 + \left(\frac{\alpha_j^P}{\pi^h}\right)^2 \sigma_{Pj}^2\right).$$

Then the probability of a C-section conditional on a patient's true medical condition $h_i$ is given by:

$$Prob\left[T_{ij} = C | h_i, \omega_j\right] = F\left(\hat{\theta}_j \left(h_i + \gamma_j\right)\right), \tag{11}$$

where $\hat{\theta}_j = \frac{1}{\sigma_{j\zeta}}$. Notice that the slope term increases with the diagnostic skill of the physician. In the special case where either there are no unobserved preferences for C-section (or unobserved medical information) then $\sigma_{Pj}^2 = 0$. In the special case where physicians disregard patient preferences (or unobserved medical information) then $\alpha_j^P$=0. In either special case, the slope is completely determined by $1/D_j$. However, even in the special case where $\alpha_j^P = 0$, the intercept term $\gamma_j$, is affected by a mix of physician beliefs, surgical skill, and prices. As discussed above Cutler et al. (2013) and Finkelstein et al. (2014) suggest that procedure choice is not generally driven by patient preferences, and hence in what follows we identify variations in the slope term as primarily reflections of diagnostic skill.

### 3.3 Measuring Physician Behavior

We now have a model that connects observed patient conditions to physician decision making. The final step is to link this behavior to observables. We cannot directly observe patient condition $h_i$. It is assumed to be uncorrelated with $\epsilon_{ij}$ in equation 7 because the contribution of each physician to $h_i^I$ is assumed to be small. If we substitute this expression into equation 9 then we can derive the probability of observing a C-section conditional on $h_i^I$.

**Proposition 2.** *The probability that physician $j$ chooses $T$=$C$ when patient condition is observed to be $h_i^I$ is given by:*

$$p_j\left(h_i^I\right) = F\left(\theta_j(h_i^I + \gamma_j)\right), \tag{12}$$

*where $\gamma_j$ defines treatment styl*e, *, and the slope term,$\theta_j$, defines the sensitivity of the doctor to the patient's condition and is given by:*

$$\theta_j = \frac{1}{\sqrt{\sigma_I^2 + \sigma_{j\zeta}^2}} \tag{13}$$

$$= \left(\sigma_I^2 + \frac{1}{D_j} + \left(\frac{B_j}{D_j} + 1\right)^2 \left(\alpha_j^P \sigma_{Pj}\right)^2\right)^{-\frac{1}{2}}, \tag{14}$$

*where $\sigma_{j\zeta}^2$ is the variance of the doctor's information conditional upon patient health, and $\sigma_I^2$ is variance of the measure of patient health given the observed birth record.*

This proposition summarizes the effects of physician characteristics on procedure choice as a function of the information that we can observe. We can directly estimate both treatment sensitivity, $\theta_j$, and treatment style, $\gamma_j$, which together define physician decision making.

Since we are measuring patient condition with error, the slope term we measure is less steep than the slope with respect to true underlying condition ($\theta_j < \frac{1}{\sigma_{j\zeta}} = \hat{\theta}_j$). Despite this issue, as long as our proxy for patient condition, $h_i^I$ is correlated with true patient condition ($\sigma_I^2$ is finite), then variations in physician characteristics will lead to variations in both the intercept, $\gamma_j$, and the slope, $\theta_j$. We now detail these effects.

### Determinants of the Intercept Term

As one can see from equation (12), any increase in $\gamma_j$, which we call treatment style to be consistent with the earlier literature, leads to an increase in the incidence of procedure C. Treatment style is affected by several attributes of physicians and their practices, as summarized in a corollary to proposition 2:

**Corollary 3.** *Treatment style and the incidence of procedure C is increasing in physician beliefs ($dp_j\left(h_j^I\right)/dh_j^0 > 0$), relative surgical skill for procedure C (($dp_j\left(h_j^I\right)/ds_j > 0$) and the relative pecuniary returns for procedure C (($dp_j\left(h_j^I\right)/dm_j > 0$). Finally, treatment style may be affected by both patient preferences and physician sensitivity to these preferences, the $\alpha_j^P \bar{h}_j^P$ term.*

All these parameters have the same basic effect upon treatment choice - they move the intercept term in (12). The effect of each of the first three parameters in corollary 3 is the same.

### Determinants of the Slope Term

The following proposition summarizes the effects of physician characteristics on the slope term.

**Corollary 4.** *The slope term, $\theta_j$, is increasing with physician diagnostic skill $\left(\frac{\partial \theta_j}{\partial D_j} > 0\right)$, decreasing with physician sensitivity $\left(\frac{\partial \theta_j}{\partial \alpha_j^p} < 0\right)$, the strength of physician prior beliefs $\left(\frac{\partial \theta_j}{\partial B_j} < 0\right)$ and with the variance of patient preferences $\left(\frac{\partial \theta_j}{\partial \sigma_{pj}^2} < 0\right)$. It is unaffected by physician skill, physician expectations, and treatment costs.*

This result follows immediately from an inspection of the formula for the slope in proposition 2. An increase in diagnostic skill means that the physician increases the weight that she places on observed patient condition, which leads to an increase in the slope. Firmer prior beliefs about the patient's condition have the opposite effect. Similarly, if the physician places more weight on patient preferences (or alternatively, on medical information that we do not observe), then this decreases the weight on the observed medical information. An increase in the variance of patient preferences can also lead to a lower weight on the observable information regarding a patient's medical condition.[5]

Consider now the relationship between diagnostic skill and the slope term, $\theta_j$. Define the elasticity of diagnostic skill with respect to $\theta_j$ as:

$$e_j^D(D_j) = \frac{D_j}{\theta_j}\frac{\partial \theta_j}{\partial D_j} > 0.$$

Using this definition and proposition 2 we have:

**Corollary 5.** *An increase in diagnostic skill increases treatment C if and only if:*

$$h_i^I \geq \hat{h}_j^I \equiv \left(1 - e_j^D(D_j)\right)\left(h_j^0 + \bar{\gamma}_j\right) - \gamma_j.$$

This result shows that diagnostic skill has an ambiguous effect on procedure choice. For patients at high risk for procedure C ($h_i^I \geq \hat{h}_j^I$), an increase in diagnostic skill increases the incidence of procedure C, while the reverse occurs for low risk patients ($h_i^I < \hat{h}_j^I$). This result is in sharp contrast to the effect of surgical skill. If a physician

[5]This result has interesting implications for patient self-selection. Suppose that there are two physicians, one who is known to have a high C-section rate, while the other has a low C-section rate. If patients know the physician's type and self select, then this will tend to reduce the variance $\sigma_{Pj}^2$ relative to the population variance of patient preferences, and thus increase the measured slope.

is better at performing a C-section then this increases the incidence of C-sections for all patients.

The contrasting effects of diagnostic and surgical skill are illustrated in Figures 1 and 2. In each figure, patients are arrayed along the X-axis from those with the lowest values of $h_i^I$ to those with the highest values. The lower line in Figure 1 illustrates the initial relationship between the observed patient condition and the probability that the intensive procedure is performed. The upper line in Figure 1 shows how this relationship would be expected to change with increases in surgical skill. The main takeaway is that one would expect an increase in the use of intensive procedures for both high and low risk patients.

Figure 2 illustrates the effect of improving diagnostic skill. From corollary 3 we have that patients with observed condition greater than $\hat{h}_j^I = -\gamma_j + \left(1 - e_j^D(D_j)\right)$ have higher C-section rates when diagnostic skill increases, and lower rates when $h_i^I$ is less than the threshold $\hat{h}_j^I$. This is illustrated in figure 2 by the move from the green/dark line to the red/light line. Thus as diagnosis improves, the use of the intensive procedure falls among those with low $h_i^I$ and increases among those with high $h_i^I$.

### 3.4 The Effect of Diagnostic and Surgical Skill on Outcomes

Let $I^C(h_i, \omega_j) = 1$ if and only if physician $j$ chooses procedure C for patient $i$ with condition $h_i$, and equal zero otherwise. Given this indicator for procedure choice, the expected medical outcome of a patient with condition $h_i$ being treated by physician $j$ is given by:

$$\begin{aligned} W(h_i, \omega_j) &= E\left\{s_j^C I^C(h_i, \omega_j) + \left(h_i + s_j^N\right)\left(1 - I^C(h_i, \omega_j)\right)\right\}, \\ &= s_j^C Prob\left[T = C | h_i, \omega_j\right] + \left(h_i + s_j^N\right) Prob\left[T = N | h_i, \omega_j\right]. \end{aligned} \tag{15}$$

However, since physicians take into account both costs, $m_j$, and patient preferences, $h_i^P$, their decisions do not maximize observed medical benefit, which complicates the computation of the effect of exogenous parameters on measured medical outcomes.

In this section we derive the effect of physician characteristics on observed medical outcome by measured risk $h_i^I$. Formally we wish to compute:

$$W\left(h_i^I, \omega_j\right) = E\left\{W(h_i, \omega_j) | h_i^I, \omega_j\right\}.$$

Since we have assumed that information about health is normally distributed, we can use results about the expectation of normally distributed random variables conditional on a truncated distribution to obtain a closed form solution for patient welfare.[6]

**Proposition 6.** *The expected medical benefit from treatment satisfies:*

$$W\left(h_i^I, \omega_j\right) = s_j^C p_j\left(h_i^I\right) + \left(s_j^N - h_i^I\right)\left(1 - p_j\left(h_i^I\right)\right) + \sigma_I^2 \frac{\partial p_j\left(h_i^I\right)}{\partial h_i^I}.$$

This is an exact formula that essentially replaces $h_i$ with $h_i^I$ plus an adjustment term $\sigma_I^2 \frac{\partial p_j\left(h_i^I\right)}{\partial h_i^I}$ to control for the fact that we do not observe $h_i$ but only an indicator, $h_i^I$. If we assume that the effect of physician characteristics on the final term in welfare, $\sigma_I^2 \frac{\partial p_j\left(h_i^I\right)}{\partial h_i^I}$, is small, then we can derive an intuitive expression for the effects of physician characteristics on outcomes.

Consider first the effect of surgical skill:

$$\frac{\partial W}{\partial s_j^C} = p_j\left(h_j^I\right) + \left(s_j + h_i^I\right)\frac{\partial p_j}{\partial s_j^C}.$$

This formula shows that the effect of skill on patient welfare can be broken into two parts. The first term is always positive, indicating that for a woman who is having the intensive procedure, more skill is always better. However, the second term is ambiguous in sign. We know that $\frac{\partial p_j}{\partial s_j^C} \geq 0$, so that other things being equal, greater doctor skill increases the probability that an intensive procedure will be performed. If $s_j + h_j^I \geq 0$, then the second term is positive and greater doctor skill enhances patient welfare. However, for a low enough value of $h_j^I$, it is possible that $s_j + h_j^I \leq 0$ (health status is in log terms, and hence is negative for low values). If $\frac{\partial p_j}{\partial s_j^C}$ is large enough, then increases in doctor skill could make patients who don't need a C-section worse off by increasing the probability that they will receive an unnecessary procedure.

Next consider the effect of physician sensitivity to patient condition, $\theta_j$. The variable is a combination of various aspects of physician characteristics, but we cannot separately observe these aspects. We do observe $\theta_j$ and $\gamma_j$ for each physician in our data, and hence can ask how outcomes would vary if we were to hold $\gamma_j$ fixed but allow $\theta_j$ to vary. Since $\theta_j$ has a first order effect on our last term, we include it, and leave out the $f''$ term. In that case we get:

$$sign \frac{\partial W}{\partial \theta_j} = sign\left\{\left(s_j + h_j^I\right)\left(h_j^I + \gamma_j\right) + \sigma_I^2\right\}.$$

This result illustrates the fact that the preferences of the physician take into account

[6]See Birnbaum (1950).

their prior beliefs, costs, and patient preferences. Hence in general $\gamma_j \neq s_j$. Whenever $h_j^I \in [min\{s_j, \gamma_j\}, max\{s_j, \gamma_j\}]$ then it is possible to have $sign\frac{\partial W}{\partial \theta_j} < 0$, but in all other cases we have a positive effect.

**Proposition 7.** *Suppose* $h_j^I \notin [min\{s_j, \gamma_j\}, max\{s_j, \gamma_j\}]$ *then increasing diagnostic skill improves medical outcomes.*

Our model suggests that average C-section rates are misleading. In particular, one can observe physicians with the same risk adjusted C-section rates but quite different outcomes: Increasing the C-section rate for the high risk and reducing it for the low risk will lead to better outcomes, even if the overall C-section rate remains constant at the population average.

# 4 Data and Methods

For our purposes of identifying the role of diagnostic skill and relating it to outcomes, C-section, which is the most common surgical procedure in the U.S., is ideal. The technology has been stable for a long time and there are detailed records on millions of births, meaning that it should be possible to use the available data to rank pregnant women in terms of their a priori risk of C-section with a fair degree of accuracy and consistency over time. Moreover, we can investigate a variety of health outcomes, including both poor outcomes for the mother and poor outcomes for the child, and thus directly relate diagnostic skill to outcomes.

The data for this project come from approximately a million Electronic Birth Certificates, (EBC) spanning 1997 to 2006, from the state of New Jersey. These records have several important features. First, in addition to information about the method of delivery, they include detailed information about the medical condition of the mother which enables us to estimate the probability of C-section for each mother. In particular, we know the mother's age, whether it is a multiple birth, whether the mother had a previous C-section, whether the baby is breech, whether there is a medical emergency such as placenta previa or eclampsia which calls for C-section delivery, and whether the mother had a variety of other risk factors for the pregnancy such as hypertension or diabetes.

Second, the birth records include detailed information about health outcomes for both the mother and the child including complications that occur during the delivery (maternal bleeding, fever, or seizures); maternal complications that occur after the delivery; fetal distress (measured by the presence of meconium); birth injuries (fracture, dislocated shoulder and other injuries); and neonatal death (death in the first 30 days of life). We also combine all of these measures into an indicator equal to one if there was "any bad outcome."

Third, the data has information about the latitude and longitude of each woman's residence, as well as codes for doctors and hospitals.[7] In our analysis, we focus on doctors and exclude midwives since only doctors can perform C-sections. Finally, the data includes demographic information about the mother such as race, education, marital status, and whether the birth was covered by Medicaid which have been shown to be related both to the probability of C-section and to birth outcomes. The inclusion of these variables may help us to control for variations in demand for C-sections by different demographic groups.

We use these data to construct analogs of the key concepts in our model. We define $F(h_i^I)$, the mother's risk of C-section, by estimating a logit model of the probability of C-section given all of the purely medical risks recorded in the birth data, as in equation (1). Since we are trying to define *medical* risk, we do not include variables such as the type of insurance and race in these logit models. The model we use is shown in column 1 of Table 1. Table 1 shows that the model predicts well, with a pseudo R-squared of almost .32.

As discussed above this model reflects actual practice, but not necessarily best practice. In principal, one might wish to estimate the model of medical risk using only the best doctors, or perhaps only the beginning of the time period when C-section rates were much lower. We have experimented with several alternative models and found that the correlation between the ranking of C-section risk produced by our model, and the ranking produced by the alternatives is above .95. These alternatives included a model with fewer risk factors, a model that used births from 1997-1999 only, and a model that used only doctors who were below the 25th percentile in terms of the fraction of births with negative outcomes in their practices. Estimates of the latter "good doctor" model are also shown in Table 1. One can see that the estimated coefficients for these "good doctors" are similar to those for all doctors suggesting that there is not a lot of controversy about the ranking of which women are the best candidates for C-section. Rather, the controversy about C-section can be interpreted as a matter of where the cutoff for C-section should occur.

Corollary 4 showed that the slope term in the model, $\theta_j$, is affected by diagnostic

[7] These codes do not identify the physician, but allow us to identify all births delivered by the same physician.We found, as a practical matter, that very few doctors practiced in more than one hospital in a single year; hence the choice of doctor also defines the choice of hospital.

skill ($D_j$). The empirical analog can be obtained for each doctor by using the estimated $\beta$'s from (1) to create the index of maternal condition $h_i^I$ (this is simply $\beta X_i$) and then estimating a regression model for each doctor's propensity to perform C-sections as a function of $h_i^I$. The estimated coefficient on $h_i^I$ , denoted by $DiagSkill_j$, is an indicator of how sensitive the doctor is to this index of observable indicators of patient risk and varies with diagnostic skill as we discussed above. The distribution of slope coefficients has a mean of 1.033 and a standard deviation of .183. The first percentile is .576 while the 99th percetile is 1.491, suggesting that doctors range from being quite insensitive to quite sensitive to maternal conditions. We normalize this measure by calculating a Z-score, for ease of interpretation.

Figure 3 plots actual C-section rates against the propensity score that we calculate. It shows that those who did not have a C-section generally had values of $F(h_i^I)$ less than .5, while those with C-sections generally had values of $F(h_i^I)$ greater than .5. However, there are individuals with no apparent risk factors who nevertheless have C-sections, and perhaps more disturbingly, there are women with many risk factors for C-section who do not receive the procedure. For a given level of medical risk, the probability of C-section increased over our sample period at all but the highest risk levels as shown in Appendix Figure 1. In fact, at the start of our sample period, New Jersey, with a rate of 24%, had a lower C-section rate than several other states, including Arkansas, Louisiana, and Mississippi, while by the end of our sample period, New Jersey had pulled ahead to have the highest C-section rate of any state, at almost 40%. Appendix Figure 2 shows that this increase was not due to a change in the underlying distribution of medical risks. The figure shows only a slight increase in the number of high risk cases, which is attributable to an increase in the number of older mothers, mothers with multiple births, and increasing numbers of women with previous C-sections (itself driven by the increasing C-section rate).

Figure 3 also shows that those who had values of $F(h_i^I)$ less than .06 (a group whom we designate the very low risk) were very unlikely to have C-sections, while those with $F(h_i^I)$ greater than .8 (a group whom we designate as the very high risk) were highly likely to have C-sections. Of the women deemed very high risk, 89% received a C-section, while among the women deemed very low risk only 6% received a C-section. 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, as noted above, 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 -.0493 (since bad outcomes are more frequent in high risk cases than in low risk cases) and a standard deviation of .0646. The first percentile of this variable is -.25, while the 99th percentile is .079. Again, we normalize this measure by calculating a Z-score for ease of interpretation.

Although relative prices for C-sections and normal deliveries have been shown to be an important determinant of C-section rates, they are not the main focus of our analysis and are not well measured in our data. We use data from the Health Care Utilization Project (HCUP), which includes hospital list charges for every discharge. For each market and year, we take the mean price of all C-section deliveries that did not involve any other procedures, less the mean price of normal deliveries without other procedures. The mean differential was $4,711 real 2006 dollars.[8]

Having constructed these measures, we estimate models of the following form:

$$Outcome_{ijt} = f(DiagSkill_j, s_j^C - s_j^N, \Delta P_{jt}, Z_{it}, month, year, zip), \tag{16}$$

where $Outcome_{ijt} \in \{0, 1\}$, where 0 is a Natural delivery (or good birth outcome) and 1 is a C-Section (or bad birth outcome), i indexes the patient, j indexes the doctor, and t indexes the year. The vector $Z_{it}$ includes maternal age (missing, less than 20, 25-34, 35 and over), education (missing, less than 12, 12, 13-15), marital status, race/ethnicity (African-American, Hispanic), and whether the birth was covered by Medicaid, as well as the child's gender and indicators for birth order. We include month and year effects in order to control for seasonal differences in outcomes and for longer term trends affecting all births in the state (e.g. due to other improvements in medical care), zip code fixed effects (3 digit) in order to control for characteristics of the location that may be associated with both medical care and outcomes, and also include indicators for missing marital status, smoking, birth order, and whether the birth occurred on a weekday. The standard errors are clustered at the level of the zip

[8] It is important to note that physician charges are generally separate from hospital charges and are not included in HCUP. Also, while Medicaid generally reimburses less than private insurance for deliveries, we do not find a significant effect of Medicaid coverage on C-section delivery, as shown in Appendix Table 1.

code in order to allow for unobserved correlations across a physician's cases.

Sample means are shown in Table 2. The estimation sample is slightly smaller than in Table 1 because while we used all births to calculate the probability of C-section, in the rest of the paper we exclude births that were not attended by a doctor, as well as those for whom we cannot calculate our measure of diagnostic skill (because there are too few births per provider).[9] These exclusions leave us with approximately 1,000 providers, who together deliver the vast majority of the babies in New Jersey over the sample period. We show sample means for all women, and for those with $F(h_i^I) \leq 0.2$ (low C-section risk) and those with $F(h_i^I) > 0.2$ (high C-section risk). This cutoff is chosen because Figure 3 suggests a gap in C-section propensities at that value, and because it divides the sample approximately in half. The first panel shows how the outcome variables vary with risk. As expected, higher risk women have more C-sections and a higher risk of a bad outcome. Examining the type of bad outcome more narrowly suggests that women at high risk of C-section are more likely to experience complications of labor and delivery as well as late maternal complications, and that their infants are at a higher risk of neonatal death.

The second panel explores the characteristics of doctors and provides some initial evidence with regard to an important question: The extent to which higher risk patients see doctors with particular characteristics. Table 2 suggests that the doctors who treat low risk patients do vary systematically from those that treat higher risk patients. As discussed above, our measures of diagnostic skill and procedural skill have been transformed into Z-scores, so in the full sample, they have a mean of zero and a standard deviation of 1. Table 2 shows that on average, high risk patients see doctors with slightly better diagnostic skills (.03 standard deviations), and slightly better surgical skills (.014 standard deviations). Conversely, low risk patients see doctors with slightly lower diagnostic skill (-.032 standard deviations) and procedural skill (-.016 standard deviations). Thus, while there is some evidence of sorting, the extent of sorting appears to be quite small. There is also some evidence that high risk patients see doctors with slightly fewer deliveries and higher shares of high risk patients in their practices. Again, however, these differences are quite small.

The third panel of the table provides an overview of selected maternal and child characteristics including race and ethnicity, maternal education, marital status, and whether the birth is covered by Medicaid. The table suggests that women at higher risk of C-section tend to be older, married, and more likely to have private insurance

[9] We also exclude a very small number of doctors who did not have at least one high risk patient and at least one low risk patient.

rather than Medicaid. They are also more likely to be delivering a first child, and are less likely to be African-American or Hispanic.

One empirical difficulty involved in estimating (16) is the possibility that women choose their doctors on the basis of their skill. If women with high risk pregnancies choose better doctors, then the estimated effect of doctor skill on birth outcomes will be biased towards zero. Table 2 suggests that there is some evidence of this type of selection, although it appears to be quite small. A second empirical problem is that we are using estimated values of diagnostic and surgical skill, which are inevitably measured with some errors.

One way to address these issues is to estimate models using market-level measures of skill as instruments for individual doctor's skill levels. Following Kessler and McClellan (1996) our definition of a hospital market is defined with reference to all of the providers actually selected by women in a particular zip code in a particular year. Specifically, we include all hospitals within ten miles of the woman's residence, plus any hospital used by more than three women from her zip code of residence in the birth year, and we consider all of the providers who practiced in those hospitals in that year as part of the relevant market. Figure 4 shows the distribution of hospitals and illustrates this way of defining markets. The figure shows that most women choose nearby hospitals, but that some women bypass nearby hospitals in favor of hospitals further away. In some cases, these are regional perinatal centers which are better equipped to deal with high risk cases. For example, women from Princeton New Jersey could give birth in the hospital in town, but many travel as far away as Morristown (two counties to the north) to deliver in other hospitals.[10] Thus, there is a distinct market, or set of provider choices, facing each woman at the time of each birth.

Given this definition of a market, we construct instruments by taking the weighted mean of the diagnostic skill and surgical skill measures for all physicians in the market in the birth year, where the weights are given by the number of deliveries by each physician.[11] We interpret this instrument as a summary measure of the choices available to a woman in a particular market. Variation in the set of providers facing each woman at a point in time comes mainly from entry and exit of providers into the various markets. By definition, these choices are affected by where women live, but recall that

[10]The figure also illustrates that the common practice of drawing a circle around a location in order to define a market is likely to be seriously misleading: A circle wide enough to include all the hospitals actually chosen would include hospitals that were never chosen, and a circle wide enough to include most hospitals could miss specialty hospitals that were further away and yet within the choice set.

[11]In the crowded northern New Jersey hospital market, we included only hospitals within five miles of the zip code centroid.

we control for zip code fixed effects in all our models. These zip code fixed effects are intended to absorb mean differences in the available health services. Hence, as long as women are not moving in order to take advantage of year to year fluctuations in the skill set of local physicians, our instruments will be valid. Table 3, which shows the first stage regressions, shows that these instruments are highly predictive.[12]

A third issue is that by construction, good diagnosticians should be less likely to perform C-sections on low risk women and more likely to perform C-sections on high risk women. Similarly, physicians with good procedural skills should have better outcomes for the high risk relative to the low risk. However, it is important to note that there is no mechanical reason for our measure of diagnostic skill to affect health outcomes, and similarly no mechanical reason for our measure of procedural skill to affect C-section rates. Thus, estimates of these two relationships form the true test of our model.

## 5 Results

Table 4 shows both Ordinary Least Squares (OLS) and Two-Stage Least Squares (TSLS) estimates of equation (16), where the dependent variable is whether there was a C-section. These models include all of the control variables discussed above. The full OLS models for the probability of C-section are shown in Appendix Table 1. Conditional on C-section risk, African-American and Hispanic women are more likely to have C-sections, as are less educated women, single women, older mothers, and mothers of first born children. These estimates suggest that the stereotype that it is primarily older, better educated white women who are "too posh to push" may be incorrect. The estimated effect of market prices is positive, but not precisely estimated.

As discussed above, the OLS coefficients on the measures of physician skill may be biased by selection and by measurement error. For example, a woman who desires a C-section regardless of her medical condition will be likely to seek a physician who does not insist on using her medical condition to determine treatment. In our taxonomy, this will be a physician with a low slope term, which we are identifying with poor diagnostic skills. In this case, OLS estimates of the coefficients on diagnostic skill will

[12] The IV estimate assumes that the instrument affects outcomes only through the quality of the doctor. Yet it is conceivable that the quality of the hospital in terms of nursing staff, for example, also matters. In this case, the IV estimate is going to pick up the "true" effect of the physician skill level, plus the nearby-hospital-specific effects. If better doctors practice in higher-quality hospitals, then the TSLS estimates could be biased upwards. In this case, the true estimate would be bounded by the OLS and IV. However, in practice we found that there was as much variation in doctor quality within hospitals as between hospitals leading us to believe that doctors are not strongly sorted into particular hospitals.

be biased towards zero. It is less clear how the coefficient on surgical skill will be affected. Other things being equal, a woman bent on surgery might prefer a better surgeon. However, diagnostic skill and surgical skill tend to be positively correlated in our data (the correlation in the raw measures is .259), so in choosing someone willing to disregard her medical condition, she may also be choosing a relatively poor surgeon, in which case the coefficient on surgical skill will also be biased downwards.

Table 4 suggests that the coefficients on both skill measures are biased towards zero in the OLS, although we do not have the precision to reject the null hypothesis that the OLS and TSLS estimates or the effects of diagnostic skill are the same. The TSLS estimates indicate that a one standard deviation increase in diagnostic skill would reduce the risk of C-section by 1.6 percentage points among women in the lower half of the risk distribution (a 15.5% reduction in the probability of C-section for these women), but would increase the probability of C-section by 1.9 percentage points (a 3.5% increase in the probability of C-section) in the upper half of the distribution. Overall, our measure of diagnostic skill has little effect, but this overall result masks the type of heterogeneity in the effects of diagnostic skill on low and high risk women that is predicted by our model.

An increase in surgical skill is estimated to increase the risk of C-section for everyone, as predicted by the model. For women in the lower half of the risk distribution, the TSLS estimate is 1.7 percentage points, indicating that a one standard deviation increase in surgical skill would increase the risk of C-section by 16.5%. Among women in the top half of the risk distribution, the increase is 3 percentage points, or 5.5%. In the case of surgical skill, the TSLS estimates are considerably larger than the OLS estimates. Table 2 does not suggest a huge amount of selection in terms of surgical skill. However, given that each surgeon has a relatively small number of very high risk and very low risk cases, and that bad outcomes are thankfully relatively rare, our measure of surgical skill is likely to be quite noisy. Hence, measurement error may account for the increase in the absolute value of the estimated coefficients when we move to TSLS.

The second panel of Table 4 shows the estimated effect of the two types of skill on the probability of any bad outcome. Once again, the OLS coefficients are smaller than the TSLS coefficients, and this is especially pronounced for the measures of surgical skill. The TSLS estimates suggest that a one standard deviation increase in diagnostic skill is associated with a 1.3 percentage point decrease in the probability of any bad outcome among both low and high risk women. This translates into a 15.3% decline among the low risk, and a 9.1% decline among the high risk. Similarly, a one standard deviation increase in surgical skill reduces the probability of any bad outcome by 42.3%

among the low risk, and by 50.3% among the high risk.

Tables 5 and 6 delve more deeply into the types of bad outcomes experienced by mothers and children, respectively. Table 5 shows the effects of skill on any bad maternal outcome, and then divides these outcomes temporally into bleeding, fever, and seizures that take place during the labor and delivery, and complications that take place after the delivery (e.g. infection or bleeding following surgery). Once again, we focus on the TSLS results which tend to be larger than the OLS estimates, especially for the surgical skill measures. Better diagnostic skill is estimated to reduce the incidence of bad maternal outcomes, especially for the low risk. Among the low risk, diagnostic skill significantly reduces the incidence of bleeding, fever, or seizures during delivery, perhaps by discouraging unnecessary surgery. Among the high risk, there is no overall effect since better diagnosis reduces the incidence of bad outcomes during delivery, but increases late maternal complications. A possible interpretation is that these women are more likely to need C-section deliveries so that providing C-section reduces the incidence of poor outcomes during delivery. However, major abdominal surgery is not without risk, and increases the probability of complications after the delivery. Better surgical skills also reduce the incidence of maternal bad outcomes, but have a greater percentage point impact among the high risk than among the low risk, which is to be expected given that the later are more likely to have surgery.

Table 6 breaks down the infant health outcomes. The first panel suggests that increases in diagnostic skill reduce poor child health outcomes, though the TSLS estimates are not very precise. The second panel indicates that there is a significant negative effect of diagnostic skill on the probability of fetal distress. This is slightly offset by a positive, though not statistically significant effect on the probability of birth injury. A possible interpretation is that infants are more likely to sustain an injury such as a dislocated shoulder if a vaginal delivery is attempted. The last panel indicates that diagnostic skill has a significant negative effect on the probability of neonatal death, but only among the high risk. This result suggests that C-section can be life-saving for infants of mothers who really require a C-section, but that unnecessary surgery does not pose a threat to the life of the infant among the low risk.

### 5.1 Robustness

Since the breakdown into high and low risk categories is arbitrary, one obvious way to explore the robustness of our results is by dividing mothers differently. Moreover, since, as we showed above, there is considerable concensus about the ranking of patients by appropriateness for C-section, we can assume that there is concensus about the high and the low risk, but perhaps controversy about the people in the middle. Table 7 shows estimates based on three risk categories, where now low risk is defined as the lowest quartile of $F(h_i^I)$ , high risk is defined as the highest quartile, and medium risk is defined as the two quartiles in the middle. The first row of Table 7 suggests that better diagnostic skill significantly reduces C-sections among the lowest risk, but has a large positive effect on the highest risk group. Better procedural skill increases C-section rates across the board.

The next panel of Table 7 indicates that the impact on "any bad outcome" is greatest for the medium and high risk groups, while procedural skill improves outcomes for all groups. Comparing the third panel of Table 7 to Table 5 indicates that better diagnostic skill has the greatest impact on preventing poor maternal outcomes among the lowest risk mothers. This is consistent with the idea that negative maternal outcomes are most likely to be caused by unnecessary surgery, since better diagnostic skill reduces unnecessary surgery among the low risk. Comparing the last panel of Table 7 to Table 6 shows that it is infants born to the highest risk mothers who benefit the most from better diagnosis in terms of preventing bad infant health outcomes. Like Table 6, this result suggests that the gravest risk to infant health occurs when women who really need a C-section do not receive one. Thus, if we only consider infant health outcomes, the trend towards higher use of C-section is not necessarily cause for alarm. It is only when we also consider maternal health that the high cost of excessive C-sections among the low risk becomes apparent.

# 6 Discussion and Conclusions

Physicians who deliver children in the United States are marvelously skilled. Their high quality surgical skills often save lives. Nevertheless, surgery is not risk free, and families must also rely on doctor's diagnostic skills. Previous work on procedure use has focused upon hospital- level rates.. Our information-based approach suggests putting the emphasis on the decision making skills of the physician. The fact that the risk adjusted C-section rates are high in a particular hospital does not imply that measures should be taken to reduce rates across the board. In some cases, C-section rates may be too low.

Suppose for example, that a C-section rate of 1/6 is desired. One way to obtain a perfect rate would be to simply roll a die and give each woman with a six a C-section. And yet we do not think that this would maximize health outcomes. Physicians in the data with flat "slopes" have both too low a C-section rate for high risk cases, and too high a C-section rate for low risk patients. Effective policies to address procedure use should consider the possibility of variation in diagnostic skill and focus on assisting physicians in making the right decisions on an individual basis. Moreover, the right decision depends on the mother-physician pair, since physicians who are more skilled at performing surgery should have higher C-section rates, other things being equal. The tools we suggest in this paper are easy to use, and can no doubt be refined. However, they demonstrate that it is possible to identify individual physicians who could benefit from decision making assistance.

Economists often suggest the use of incentive systems to improve performance. In the case of high C-section rates this would mean reducing the pecuniary incentive to do the procedure, either by decreasing the price of a C-section, or increasing the price of a natural delivery. Our results do not support such a policy because optimal procedure use is a complex function of both physician and patient characteristics.

The previous literature on treatment choice emphasizes that it is affected by physician skill, but only allows physician skill to vary along a single dimension which can be thought of as technical skill in executing procedures, or surgical skill. Taking a cue from the literature on expert decision making we develop a model that includes an additional dimension of skill: Diagnostic decision making. In our model, a good doctor is one who is not only technically skilled, but who is also able to draw the correct inferences from the available data in order to match patients correctly to the procedures that are most likely to benefit them.

This simple framework yields rich predictions and allows us to distinguish between two factors which we identify with the two types of skill. The model shows that better procedural skill leads to higher use of intensive procedures across the board, for both high and low risk patients. In contrast, better diagnostic skill results in fewer procedures for the low risk, but more procedures for the high risk. That is, better diagnostic skill improves the matching between patients and procedures and thus leads to better health outcomes in both groups.

We provide an application of our model using data on C-sections, the most common surgical procedure performed in the U.S.. We show that improving diagnostic skill by one standard deviation would reduce C-section rates by 15.5% in the lower half of the distribution of C-section risk, but would actually increase C-sections by 5.5% in the top half of the distribution. This finding suggests that not only are there two many C-sections among women without risk factors, but there are too few C-sections in the group who really needs them.

Our work highlights the importance of diagnostic decision making in medicine and suggests an empirical approach to measuring it: Given a prediction of a patient's medical appropriateness for a procedure, a doctor's diagnostic skill can be evaluated by looking at whether they are responsive to this information. We show that doctors who are not responsive to the publicly observed patient medical information typically achieve worse health outcomes. This finding suggests then, that the medical information contained in sources such as electronic patient records could be used to improve medical decision making. We are not suggesting that doctors be replaced by machines, but that a doctor's individual expertise, which perforce depends on his or her individual experience, could be enhanced by applying simple algorithms to the "big data" contained in millions of administrative medical records.

## References


Arlen, J. and W. B. MacLeod (2005, Fall). Torts, expertise, and authority: Liability of physicians and managed care organizations. *Rand Journal of Economics 36*(3), 494–519.

Baicker, K., E. S. Fisher, and A. Chandra (2007, MAY-JUN). Malpractice liability costs and the practice of medicine in the medicare program. *Health Affairs 26*(3), 841–852.

Baker, G.R. amd MacIntosh-Murray, A., C. Porcellato, K. Dionne, L. amd Stelmacovich, and K. Born (Eds.) (2008). *High Performing Healthcare Systems: Delivering Quality by Design.* Toronto: Longwoods Publishing.

Birnbaum, Z. W. (1950). Effect of linear truncation on a multinormal population. *The Annals of Mathematical Statistics 21*(2), 272–279.

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.

Commission, T. J. (2014). Specification manual for joint commission national quality core measures. Technical Report Version 2014A1, The Joint Commission. URL: https://manual.jointcommission.org/releases/TJC2014A1/.

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.

Cutler, D., J. Skinner, A. D. Stern, and D. Wennberg (2013, August). Physician beliefs and patient preferences: A new look at regional variation in health care spending. Technical Report 19320, NBER.

DeGroot, M. H. (1972). *Optimal Statistical Decisions*. New York, NY: McGraw-Hill Book C.

Doi, K. (2007, Jun). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Comput Med Imaging Graph 31*(4), 198–211.

Dranove, D., S. Ramanarayanan, and A. Sfekas (2011). Does the market punish aggressive experts? Evidence from cesarean sections. *B E Journal oOf Economic Analysis & Policy 11*(2).

Dubay, L., R. Kaestner, and T. Waidmann (1999, August). The impact of malpractice fears on cesarean section rates. *Journal of Health Economics 18*(4), 491–522.

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.

Finkelstein, A., M. Gentzkow, and H. Williams (2014, July). Sources of geographic variation in health care: Evidence from patient migration. Technical report, University of Chicago.

Garber, A. M. and J. Skinner (2008). Is american health care uniquely inefficient? *The Journal of Economic Perspectives 22*(4), pp. 27–50.

Gawande, A. (2009). *The Checklist Manifesto: Getting Things Right*. New York: Picador.

Gaynor, M., J. B. Rebitzer, and L. J. Taylor (2004, Aug). Physician incentives in health maintenance organizations. *Journal of Political Economy 112*(4), 915–931. English.

Grove, W., D. H. Zald, B. S. Lebow, B. E. Snitz, and C. Nelson (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment 12*, 19–30.

Gruber, J., J. Kim, and D. Mayzlin (1999, AUG). Physician fees and procedure intensity: the case of cesarean delivery. *Journal of Health Economics 18*(4), 473–490.

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

Johnson, E. M. and M. M. Rehai (2014, January). Physicians treating physicians: Information and incentives in childbirth. mimeo, University of British Columbia, Vancouver, BC.

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

Kessler, D. and M. McClellan (1996). Do doctors practice defensive medicine? *Quarterly Journal of Economics 111*(2), 353–90.

Kozhimannil, K. B., M. R. Law, and B. A. Virnig (2013). Cesarean delivery rates vary tenfold among us hospitals; reducing variation may address quality and cost issues. *Health Affairs 32*(3), 527–535.

McCourt, C., J. Weaver, H. Statham, S. Beake, J. Gamble, and D. K. Creedy (2007). Elective cesarean section and decision making: A critical review of the literature. *Birth 34*(1), 65–79.

Meehl, P. E. (1954). *Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press.

Smith, G., M. Dellens, I. White, and J. Pell (2004, DEC). Combined logistic and bayesian modeling of cesarean section risk. *American Journal of Obstetrics and Gynecology 191*(6), 2029–2034.

Srinivas, S. K., C. Fager, and S. A. Lorch (2010). Evaluating risk-adjusted cesarean delivery rate as a measure of obstetric quality. *Obstetrics and Gynecology 115*(5), 1007–1013.

# 7 Figures

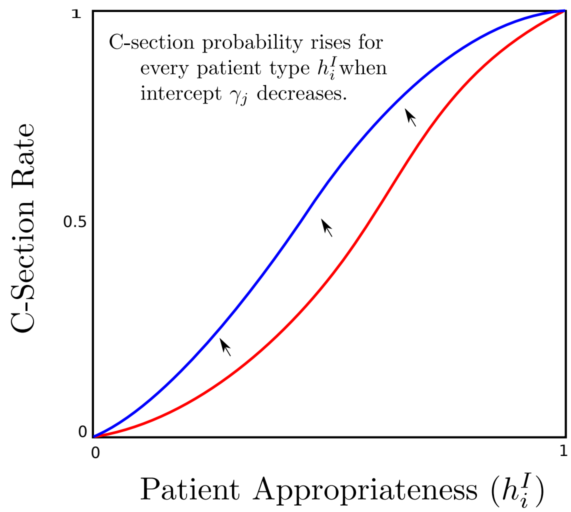


Figure 1: Effect of Intercept upon Procedure Use

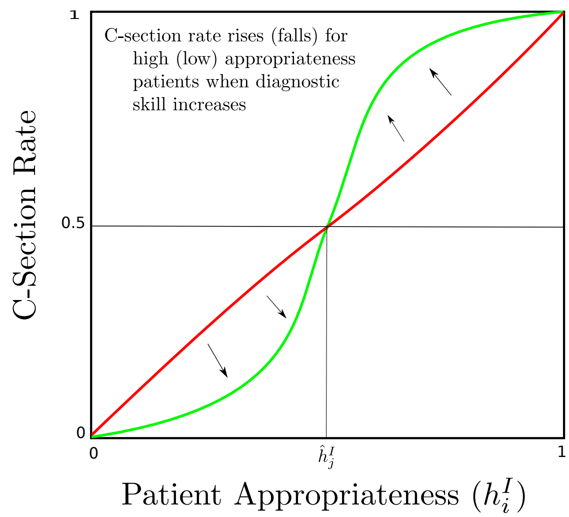


Figure 2: The Effect of Diagnostic Skill on Procedure Choice

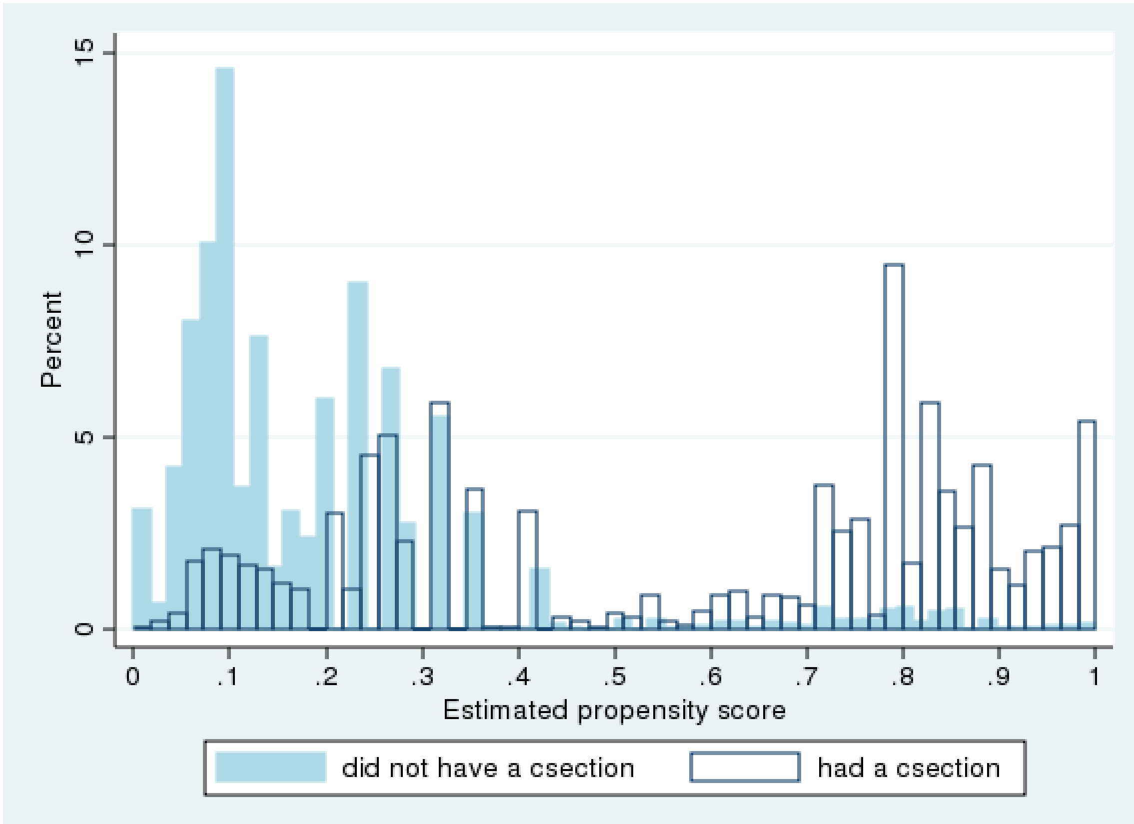


Figure 3: Predicting C-sections Using the Logit Model

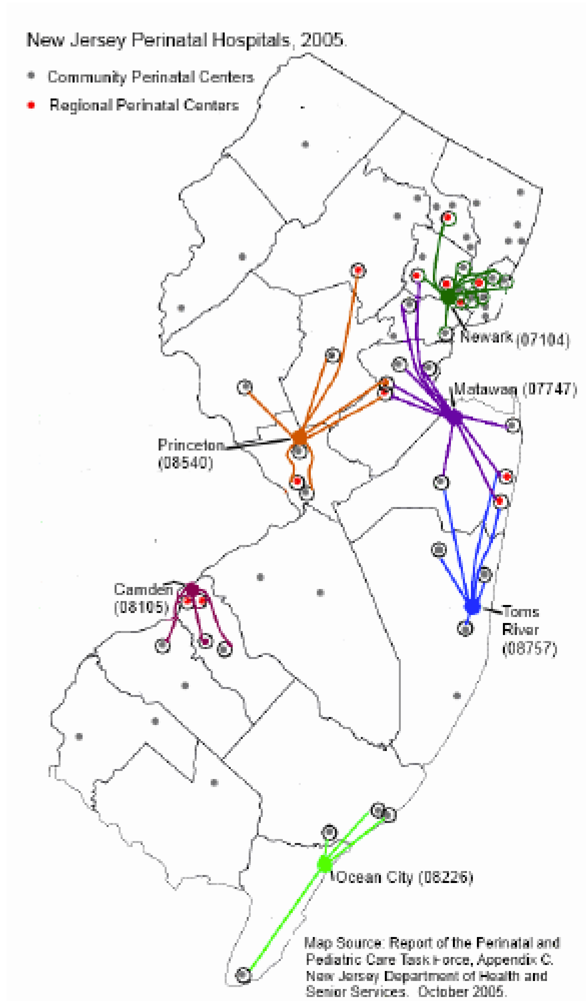


Figure 4: Illustrating the Definition of a Market

# 8 Tables

**Table 1: Logistic Regression Model of C-section Risk (rho)**

| | 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 2: Means for Full Sample and by Probability of C-Section**

| C-section Risk: | Full Sample | Low Risk of C-Section | High Risk of C-section |
|---|---|---|---|
| Outcomes | | | |
| C-Section Rate | 0.331 | 0.103 | 0.545 |
| Any Bad Outcome | 0.127 | 0.111 | 0.143 |
| Bad Maternal Outcome | 0.055 | 0.037 | 0.073 |
| Bleeding, Fever, Seizures at Delivery | 0.039 | 0.024 | 0.053 |
| Late Maternal Complications | 0.019 | 0.014 | 0.024 |
| Bad Child Outcome | 0.080 | 0.080 | 0.081 |
| Fetal Distress | 0.071 | 0.073 | 0.069 |
| Birth Injury | 0.003 | 0.003 | 0.003 |
| Neonatal death | 0.004 | 0.003 | 0.006 |
| Doctor Characteristics | | | |
| # Deliveries per doctor | 1019.45 | 1030.34 | 1009.22 |
| | (650.15) | (674.73) | (626.00) |
| Diagnostic Skill | 0.000 | -0.032 | 0.030 |
| | (1.000) | (1.013) | (0.987) |
| Procedural Skill Differential | 0.000 | -0.016 | 0.014 |
| | (1.000) | (1.034) | (0.966) |
| Market Price Differential ($1000) | 4.711 | 4.687 | 4.734 |
| | (1.606) | (1.590) | (1.621) |
| Share High Risk | 0.122 | 0.116 | 0.127 |
| | | | |
| Mother & Child Characteristics | | | |
| African American | 0.158 | 0.185 | 0.132 |
| Hispanic | 0.210 | 0.388 | 0.179 |
| Married | 0.713 | 0.645 | 0.776 |
| High School Dropout | 0.128 | 0.177 | 0.082 |
| Teen mom | 0.030 | 0.052 | 0.009 |
| Mom Age 35 or More | 0.238 | 0.221 | 0.254 |
| Smoked | 0.081 | 0.090 | 0.073 |
| Child Male | 0.513 | 0.514 | 0.513 |
| Child First Born | 0.398 | 0.200 | 0.584 |
| Medicaid | 0.206 | 0.260 | 0.155 |
| # of Observations | 968748 | 469170 | 499578 |

Notes: The analysis sample excludes birth attendants who were not physicians, and birth attendants who had too few deliveries for a measure of diagnositic skill to be computed. Standard deviations in parentheses.

**Table 3: First Stage Regressions of Doctor Skill Measures on Market Skill Measures**

| | Doctor Diagnostic Skill | | | Doctor Surgical Skill | | |
|---|---|---|---|---|---|---|
| | All | Low | High | All | Low | High |
| Market Diagnosis | 0.353 | 0.356 | 0.347 | -0.026 | -0.024 | -0.028 |
| | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Market Surgical | -0.014 | -0.009 | -0.019 | 0.284 | 0.29 | 0.276 |
| | (0.001) | (0.002) | (0.002) | (0.002) | (0.003) | (0.003) |
| R-squared | 0.165 | 0.179 | 0.152 | 0.098 | 0.105 | 0.09 |

Notes: Standard errors clustered at the 3-digit zip code level. Regressions also include market price, estimated C-section risk, indicators for African-American, Hispanic, 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.

**Table 4: Effect of Doctor Diagnostic 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** | | | | | | |
| Diagnostic Skill | 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 | 0.03 |
| | (0.002) | (0.001) | (0.002) | (0.010) | (0.008) | (0.011) |
| R-sq/Chi-sq. | 0.41 | 0.044 | 0.319 | 230000 | 12674 | 88123 |
| **Dep. Var: Any Bad Outcome** | | | | | | |
| Diagnostic Skill | -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. | 0.02 | 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, Hispanic, 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.

**Table 5: Effect of Doctor Diagnostic and Surgical Skill on Maternal Health Outcomes**

| C-section Risk: | OLS All | OLS Low | OLS High | TSLS All | TSLS Low | TSLS High |
|---|---|---|---|---|---|---|
| **Dep. Var: Any Bad Maternal Outcome** | | | | | | |
| Diagnostic Skill | -0.005 | -0.004 | -0.005 | -0.004 | -0.005 | -0.003 |
| | (0.001) | (0.001) | (0.001) | (0.003) | (0.002) | (0.003) |
| Procedural Skill Difference | -0.013 | -0.005 | -0.022 | -0.035 | -0.023 | -0.049 |
| | (0.002) | (0.001) | (0.002) | (0.007) | (0.007) | (0.008) |
| R-sq/Chi-sq. | 0.018 | 0.013 | 0.016 | 4267 | 10269 | 1988 |
| **Dep. Var: Bleeding, Fever, Seizures During Delivery** | | | | | | |
| Diagnostic Skill | -0.006 | -0.004 | -0.008 | -0.012 | -0.008 | -0.016 |
| | (0.000) | (0.000) | (0.001) | (0.002) | (0.001) | (0.003) |
| Procedural Skill Difference | -0.007 | -0.001 | -0.013 | -0.009 | -0.004 | -0.018 |
| | (0.001) | (0.000) | (0.001) | (0.003) | (0.002) | (0.004) |
| R-sq/Chi-sq. | 0.013 | 0.009 | 0.011 | 7007 | 3465 | 2340 |
| **Dep. Var: Maternal Complications After Delivery** | | | | | | |
| Diagnostic Skill | 0.001 | -0.0001 | 0.002 | 0.008 | 0.003 | 0.013 |
| | (0.001) | (0.001) | (0.001) | (0.002) | (0.002) | (0.003) |
| Procedural Skill Difference | -0.007 | -0.004 | -0.011 | -0.028 | -0.021 | -0.036 |
| | (0.002) | (0.001) | (0.002) | (0.006) | (0.006) | (0.007) |
| R-sq/Chi-sq. | 0.017 | 0.013 | 0.02 | 25060 | 997 | 645 |
| # 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.

**Table 6: Effect of Doctor Diagnostic and Surgical Skill on Child Health Outcomes**

| C-section Risk: | OLS All | OLS Low | OLS High | TSLS All | TSLS Low | TSLS High |
|---|---|---|---|---|---|---|
| **Dep. Var: Any Bad Infant Outcome** | | | | | | |
| Diagnostic Skill | -0.005 | -0.005 | -0.006 | -0.010 | -0.009 | -0.010 |
| | (0.001) | (0.001) | (0.001) | (0.007) | (0.007) | (0.007) |
| Procedural Skill Difference | -0.006 | -0.004 | -0.008 | -0.031 | -0.029 | -0.032 |
| | (0.001) | (0.001) | (0.002) | (0.009) | (0.009) | (0.009) |
| R-sq/Chi-sq. | 0.013 | 0.01 | 0.017 | 16421 | 1108 | 2099 |
| | | | | | | |
| **Dep. Var: Fetal Distress** | | | | | | |
| Diagnostic Skill | -0.003 | -0.004 | -0.003 | -0.012 | -0.012 | -0.012 |
| | (0.001) | (0.001) | (0.001) | (0.006) | (0.006) | (0.006) |
| Procedural Skill Difference | -0.007 | -0.001 | -0.013 | -0.024 | -0.025 | -0.023 |
| | (0.001) | (0.000) | (0.001) | (0.003) | (0.002) | (0.004) |
| R-sq/Chi-sq. | 0.013 | 0.009 | 0.011 | 7007 | 3465 | 2340 |
| | | | | | | |
| **Dep. Var: Birth Injury** | | | | | | |
| Diagnostic Skill | 0.0001 | 0.0001 | 0.0001 | 0.004 | 0.003 | 0.005 |
| | (0.000) | (0.000) | (0.000) | (0.003) | (0.002) | (0.004) |
| Procedural Skill Difference | -0.001 | -0.001 | -0.002 | -0.009 | -0.006 | -0.011 |
| | (0.001) | (0.001) | (0.001) | (0.004) | (0.003) | (0.006) |
| R-sq/Chi-sq. | 0.003 | 0.002 | 0.004 | 268 | 392 | 563 |
| | | | | | | |
| **Dep. Var: Neonatal Death** | | | | | | |
| Diagnostic Skill | -0.002 | -0.001 | -0.002 | -0.001 | -0.0003 | -0.002 |
| | (0.000) | (0.000) | (0.000) | (0.001) | (0.000) | (0.001) |
| Procedural Skill Difference | -0.001 | -0.0003 | -0.002 | 0.001 | 0.0006 | 0.001 |
| | (0.000) | (0.000) | (0.000) | (0.001) | (0.000) | (0.001) |
| R-sq/Chi-sq. | 0.007 | 0.004 | 0.01 | 2427 | 1445 | 2026 |
| # 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.

**Table 7: TSLS Estimates of Effect Diagnostic and Surgical Skill, Three Risk Categories**

| C-section Risk: | Low p(csect)<.084 | Medium p(csect)>=.084 p(csect)<=.439 | High p(csect)>.439 |
|---|---|---|---|
| **Dep. Var: C-section** | | | |
| Diagnostic Skill | -0.015 | -0.013 | 0.044 |
| | (0.004) | (0.009) | (0.006) |
| Procedural Skill Difference | 0.014 | 0.022 | 0.038 |
| | (0.007) | (0.012) | (0.012) |
| Chi-sq. | 5208 | 29616 | 28375 |
| | | | |
| **Dep. Var: Any Bad Outcome** | | | |
| Diagnostic Skill | -0.009 | -0.018 | -0.010 |
| | (0.007) | (0.008) | (0.003) |
| Procedural Skill Difference | -0.043 | -0.058 | -0.078 |
| | (0.006) | (0.008) | (0.005) |
| Chi-sq. | 5131 | 17881 | 4699 |
| | | | |
| **Dep. Var: Bad Maternal Outcome** | | | |
| Diagnostic Skill | -0.044 | -0.008 | 0.003 |
| | (0.002) | (0.004) | (0.004) |
| Procedural Skill Difference | -0.017 | -0.033 | -0.06 |
| | (0.006) | (0.009) | (0.008) |
| Chi-sq. | 609 | 2209 | 3330 |
| | | | |
| **Dep. Var: Bad Infant Outcome** | | | |
| Diagnostic Skill | -0.006 | -0.011 | -0.013 |
| | (0.006) | (0.010) | (0.004) |
| Procedural Skill Difference | -0.029 | -0.034 | -0.025 |
| | (0.007) | (0.011) | (0.007) |
| Chi-sq. | 19209 | 3809 | 3997 |
| # Observations | 251965 | 473011 | 243869 |

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.

**Table 8: TSLS Estimates of Effect Diagnostic and Surgical Skill, Three Risk Categories First Births Only**

| C-section Risk: | Low<br>p(csect)<.217 | Medium<br>p(csect)>=.217<br>p(csect)<=.309 | High<br>p(csect)>.309 |
|---|---|---|---|
| **Dep. Var: C-section** | | | |
| Diagnostic Skill | -0.018 | -0.015 | 0.003 |
| | (0.007) | (0.010) | (0.014) |
| Procedural Skill Difference | 0.021 | 0.022 | 0.028 |
| | (0.013) | (0.012) | (0.017) |
| Chi-sq. | 4056 | 4878 | 82795 |
| | | | |
| **Dep. Var: Any Bad Outcome** | | | |
| Diagnostic Skill | -0.025 | -0.02 | 0.000 |
| | (0.007) | (0.011) | (0.008) |
| Procedural Skill Difference | -0.066 | -0.067 | -0.084 |
| | (0.011) | (0.010) | (0.009) |
| Chi-sq. | 4569 | 18187 | |
| | | | |
| **Dep. Var: Bad Maternal Outcome** | | | |
| Diagnostic Skill | -0.005 | -0.011 | 0.001 |
| | (0.005) | (0.004) | (0.004) |
| Procedural Skill Difference | -0.043 | -0.039 | -0.054 |
| | (0.015) | (0.009) | (0.010) |
| Chi-sq. | 1152 | 6165 | 323 |
| | | | |
| **Dep. Var: Bad Infant Outcome** | | | |
| Diagnostic Skill | -0.022 | -0.009 | 0.0004 |
| | (0.006) | (0.013) | (0.009) |
| Procedural Skill Difference | -0.032 | -0.04 | -0.045 |
| | (0.009) | (0.013) | (0.010) |
| Chi-sq. | 1840 | 1359 | 690 |
| # Observations | 95123 | 184238 | 105752 |

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.

# A Appendix - Proofs

## Proof of Proposition 1

*Proof.* If $x \sim N\left(m, \sigma^2\right)$ has a normal prior distribution, and one has an observation $y = x + \epsilon$, where $\epsilon \sim N\left(0, \sigma_y^2\right)$ is normally distributed and independent of $x$, then from Theorem 1, DeGroot (1972), section 9.5, the posterior distribution of $x \sim N\left(\pi m + (1-\pi) y, \rho_x + \rho_y\right)$, were $\rho_x = \frac{1}{\sigma^2}$ and $\rho_y = \frac{1}{\sigma_y^2}$ are the precisions of x and y, while $\pi = \frac{\rho_x}{\rho_x + \rho_y}$ is the weight on prior mean.

The normal distribution is called a conjugate family because when the prior and signals are normally distributed, then so is the posterior. This allows for very simple linear learning rules. We can use other distributions, but it would greatly complicate the analysis while providing few benefits in terms of new insights. □

## Proof of Proposition 2

*Proof.* From 9 we have $T = C$ iff:

$$h_i^I + \frac{1}{\pi^h}\left(\pi^0 h_j^0 + s_j + m_j + \alpha_j^P \bar{h}_j^P\right) \geq -\left(\epsilon_i^I + \epsilon_{ij}^h\right) - \frac{\alpha_j^P \epsilon_{ij}^P}{\pi^h}. \tag{17}$$

The right hand side is a normal distribution with zero mean and variance:

$$\sigma_j^2 = \left(\sigma_I^2 + \frac{1}{D_j} + \left(\frac{\alpha_j^s \sigma_P}{\pi^h}\right)^2\right) \tag{18}$$

Hence, we can write (17) as:

$$\frac{1}{\sigma_j}\left(h_i^I + \frac{1}{\pi^h}\left(\pi^0 h_j^0 + s_j + m_j + \alpha_j^P \bar{h}_j^P\right)\right) \geq \epsilon,$$

where $\epsilon \sim N(0,1)$. Hence we have:

$$p_j\left(h_i^I\right) = F\left(\frac{1}{\sigma_j}\left(h_i^I + \frac{1}{\pi^h}\left(\pi^0 h_j^0 + s_j + m_j + \alpha_j^P \bar{h}_j^P\right)\right)\right),$$

from which we obtain the result. □

## Proof of proposition 6

*Proof.* We can write welfare as:

$$
\begin{aligned}
W(h_i,\omega_j) &= s_j^C Prob\left[T_{ij}=C|h_i,\omega_j\right] + E\left\{-h_i + s_j^N|T_{ij}=N,h_i,\omega_j\right\} Prob\left[T_{ij}=N|h_i,\omega_j\right] \\
&= s_j^C Prob\left[T_{ij}=C|h_i,\omega_j\right] + s_j^N Prob\left[T_{ij}=N|h_i,\omega_j\right] \\
&- E\left\{h_i|T_{ij}=N,h_i,\omega_j\right\} Prob\left[T_{ij}=N|h_i,\omega_j\right].
\end{aligned}
$$

Next we condition on $h_i^I$ and observe that $E\left\{E\left\{X|h_i,\omega_j\right\}|h_i^I,\omega_j\right\} = E\left\{X|h_i^I,\omega_j\right\}$, since this is strictly less information. First, we already have from equation 12:

$$
Prob\left[T_{ij}=C|h_i^I,\omega_j\right] = p_j\left(h_i^I\right).
$$

Next we have from 9:

$$
\begin{aligned}
E\left\{h_i|T_{ij}=N,h_i^I,\omega_j\right\} &= E\left\{h_i^I - \epsilon_i^I|h_i^I - \epsilon_i^I + \gamma_j \le \zeta_{ij}\right\} \\
&= E\left\{h_i^I - \epsilon_i^I|h_i^I + \gamma_j \le \zeta_{ij} + \epsilon_i^I\right\}.
\end{aligned}
$$

From Birnbaum (1950) we have that if $X$ and $Z$ are two normally distributed random variables with variances $\sigma_X^2$ and $\sigma_Y^2$ then:

$$
E\left\{X|q \le Z\right\} = E\left\{X\right\} + \frac{cov\left(X,Z\right)}{\sigma_Q} R\left(\frac{q - E\left\{Z\right\}}{\sigma_Q}\right),
$$

where $R(x) = \frac{f(x)}{1-F(x)}$ is the Mills ratio for the Normal distribution. Applying this formula with $X = h_i^I - \epsilon_i^I$, $Z = \zeta_{ij} + \epsilon_i^I$ and $q = h_i^I + \bar{\gamma}_j$ we get:

$$
E\left\{h_i|T_{ij}=N,h_i^I,\omega_j\right\} = h_i^I - \frac{\sigma_I^2}{\sigma_j} R\left(\frac{h_i^I + \gamma_j}{\sigma_j}\right),
$$

where $\sigma_j$ is defined in 18. Notice that $\theta_j = \frac{1}{\sigma_j}$ and $p_j\left(h_i^I\right) = F\left(\theta_j\left(h_i^I + \gamma_j\right)\right)$.

Thus we get:

$$
\begin{aligned}
W\left(h_i^I,\omega_j\right) &= E\left(W\left(h_i,\omega_j\right)\right) \\
&= s_j^C p_j\left(h_i^I\right) + s_j^N\left(1 - p_j\left(h_i^I\right)\right) \\
&- \left(h_i^I - \sigma_I^2\theta_j R\left(\theta_j\left(h_i^I + \gamma_j\right)\right)\right)\left(1 - p_j\left(h_i^I\right)\right) \\
&= s_j^C p_j\left(h_i^I\right) + \left(s_j^N - h_j^I\right)\left(1 - p_j\left(h_i^I\right)\right) \\
&+ \sigma_I^2\theta_j f\left(\theta_j\left(h_i^I + \gamma_j\right)\right).
\end{aligned}
$$

Now $\frac{\partial F\left(\theta_j\left(h_i^I+\gamma_j\right)\right)}{\partial h_i^I} = \theta_j f\left(\theta_j\left(h_i^I+\gamma_j\right)\right)$ and therefore we may write:

$$W\left(h_i^I, \omega_j\right) = s_j^C p_j\left(h_i^I\right) + \left(s_j^N - h_j^I\right)\left(1 - p_j\left(h_i^I\right)\right) + \sigma_I^2 \frac{\partial p_j\left(h_i^I\right)}{\partial h_i^I}.$$

$\square$

# B Appendix - Table

**Appendix Table 1: Effect of Diagnostic and Surgical Skill on Probability of C-Section**
**Ordinary Least Squares**

| C-section Risk: | All | Low | High |
|---|---|---|---|
| Diagnostic Skill | 0.004 | -0.011 | 0.019 |
| | (0.002) | (0.002) | (0.002) |
| Procedural Skill Difference | 0.003 | 0.003 | 0.003 |
| | (0.002) | (0.001) | (0.002) |
| Market Price (coeff x 100) | 0.276 | 0.291 | 0.285 |
| | (0.226) | (0.249) | (0.221) |
| C-section Risk | 1.002 | 0.902 | 0.906 |
| | (0.007) | (0.069) | (0.009) |
| African-American | 0.05 | 0.047 | 0.050 |
| | (0.004) | (0.003) | (0.005) |
| Hispanic | 0.036 | 0.024 | 0.051 |
| | (0.003) | (0.002) | (0.005) |
| Less than High School | 0.022 | 0.019 | 0.026 |
| | (0.003) | (0.002) | (0.005) |
| High School | 0.026 | 0.022 | 0.032 |
| | (0.001) | (0.002) | (0.003) |
| Some College | 0.012 | 0.011 | 0.013 |
| | (0.001) | (0.002) | (0.002) |
| Married | -0.007 | -0.009 | 0.006 |
| | (0.002) | (0.003) | (0.003) |
| Medicaid | 0.005 | 0.007 | 0.001 |
| | (0.004) | (0.004) | (0.006) |
| Teen Mom | -0.013 | -0.023 | 0.012 |
| | (0.004) | (0.005) | (0.009) |
| Mother 25-34 | 0.019 | 0.028 | 0.005 |
| | (0.003) | (0.002) | (0.004) |
| Mother 35+ | 0.025 | 0.041 | 0.013 |
| | (0.003) | (0.003) | (0.005) |
| Mother Smoked | 0.007 | 0.010 | 0.004 |
| | (0.004) | (0.003) | (0.006) |
| Child Male | 0.023 | 0.018 | 0.027 |
| | (0.001) | (0.001) | (0.002) |
| Child 2nd Born | -0.013 | -0.040 | 0.051 |
| | (0.003) | (0.008) | (0.004) |
| Child 3rd Born | -0.018 | -0.043 | 0.032 |
| | (0.003) | (0.009) | (0.006) |
| Child 4th Born or Higher | -0.022 | -0.034 | 0.001 |
| | (0.006) | (0.010) | (0.010) |
| R-squared | 0.41 | 0.044 | 0.319 |
| # Observations | 968845 | 469204 | 499641 |

Notes: Standard errors clustered by 3 digit zip code. Regressions also include indicators for month and year of birth and 3-digit zip code, as well as indicators for missing education, marital status, Medicaid coverage, smoking, parity, and an indicator for birth on a weekday.

# C Appendix - Figures

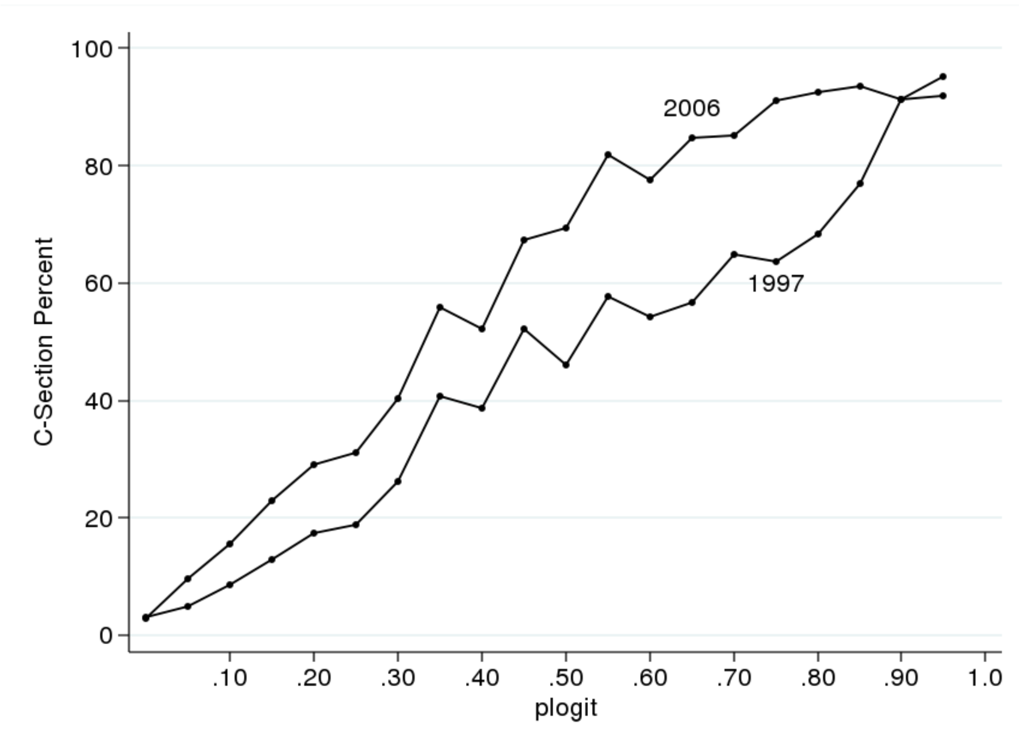


Figure 1: Shift in Probability of C-section Given Medical Risk Over Time

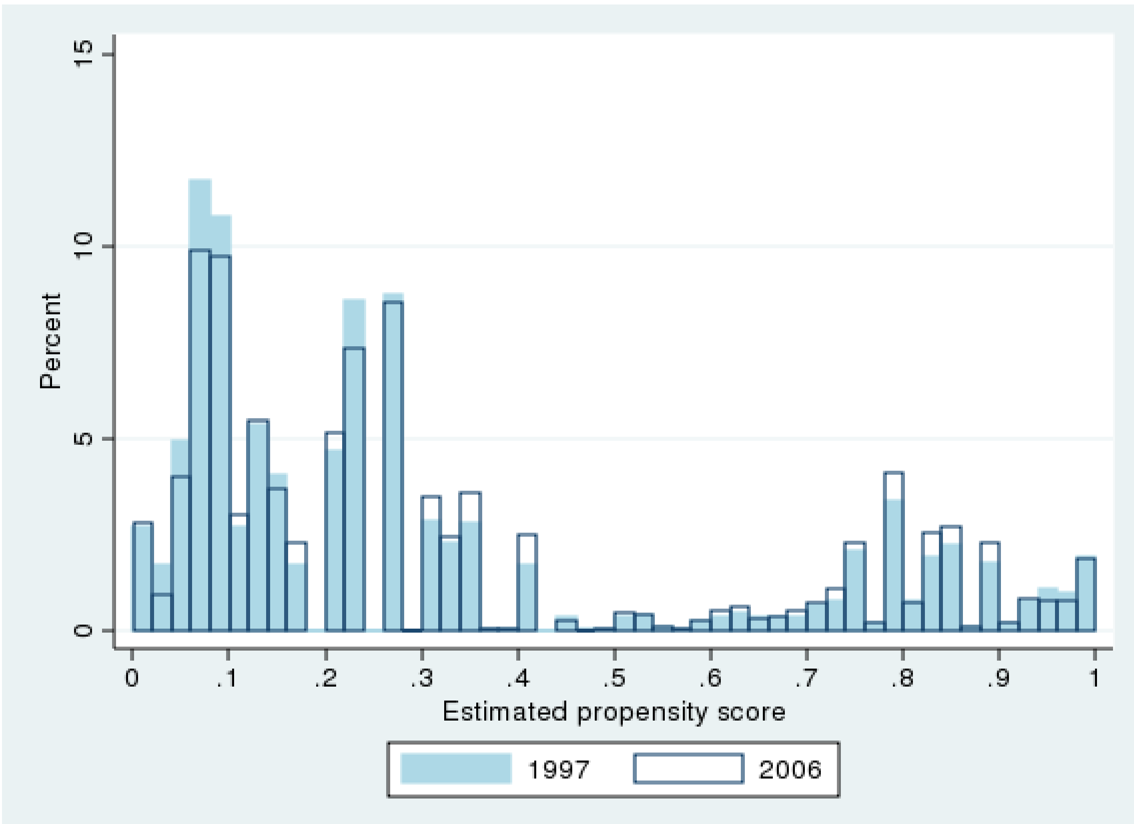


Figure 2: Shift in Medical Risks over Time