Beyond Personality Traits and Financial Incentives: Bias and Variation in Medical Practices as a Systematic Result of Experiential Learning

Navid Ghaffarzadegan

Department of Public Administration and Policy
Milne 300, Rockefeller College of Public Affairs and Policy
University at Albany, State University of New York

navidg@gmail.com

Abstract

There are several indicators of abundant sub-optimal decisions in medicine. Two of the common ones are overuse of defensive medical practices such as medical tests (bias toward more tests), and variation of medical diagnoses and treatments for medically similar patients (practice variation). Besides patients' characteristics and preferences, and the regional characteristics (such as culture), the most common explanations for practice variation and bias relate to physicians' personality traits (e.g., the level of risk aversion) and their financial incentives. We develop a theory that offers a new explanation for variation and bias in practice. With the help of a simulation model, we show that practice variation and bias do not have to be caused by personality traits and financial incentives, but can endogenously emerge through daily practices and outcome learning even among physicians with similar training working in the same region. In other words, the characteristics of medical tasks themselves can result in practice variation and bias. Specifically, a physician's exposure to outcome feedback, a physician's ability to evaluate different forms of practice, and a physician's accumulated experience with a given approach all contribute to practice variation and bias. A preliminary validation of the results is achieved by comparing projected results with actual data from cesarean section surgery in the states of New York and Florida.

Keywords: Experiential Learning, Medical Decision Making, Practice Variation, Outcome feedback, Conditional Feedback

1. INTRODUCTION

There are several indicators of abundant sub–optimal decisions in medicine. The Institute of Medicine (2000) estimates that 98,000 people die in the United States hospitals every year as a result of preventable mistakes, and the Dartmouth Atlas of Health Care argues that decision making factors are central in health disparities in the country (Fisher, Bynum, & Skinner, 2009). Sub-optimal medical decisions, in aggregate, contribute to an inefficient healthcare system, a major concern in the current US healthcare reform.

Two of the most common indicators for sub–optimal practices in medicine are practice variation and over–utilization bias (Fisher et al., 2009; Institute Of Medicine, 2000, 2003; Wennberg, Fisher, & Skinner, 2002; Wennberg, Freeman, & Culp, 1987; Wennberg & Gittelsohn, 1973). First, it is shown that different doctors do not make similar decisions for medically similar patients and in many cases they disagree. For example, controlling for patients' health risks, different obstetricians have different rates of c-section surgeries (Epstein & Nicholson, 2009). Similar patterns of disagreement across different physicians happen in prescribing cancer diagnostic tests and treatments (Bynum, Song, & Fisher, 2010), pediatric services (Sorum et al., 2002), and psychiatric services (Way, Allen, Mumpower, Stewart, & Banks, 1998). Practice variation for medically similar patients has been argued to be an indicator of sub-optimal healthcare system (Fisher et al., 2009).

Second, on average physicians prescribe more tests than needed, and they incorporate more surgeries than necessary (bias toward over utilization of resources). The current discussion around the optimal frequency of mammography is an example of when a standard of efficiency (as determined by an expert panel) may not be used in practice (Welch, 2010). Similar arguments have been made for the excessive frequency of other medical tests and the general over–utilization of medical resources (Bynum et al., 2010).

Regional characteristics have been argued to produce variation in medical expenditures (Fisher et al., 2009; Sutherland, Fisher, & Skinner, 2009). The Dartmouth atlas of healthcare offers a lot of evidence that in some regions there is more health expenditure than others. In these studies, they find that only 30% of the excess spending in the highest cost regions can be related to income and health and the rest are regional factors. In other words, some regions have a higher level of health expenditure due to the cultural, demographic and industrial factors. Such a higher level of expenditure does not necessarily result in a better outcome (quality of healthcare services), and therefore is an indicator for an inefficient healthcare system (Sutherland et al., 2009). Although they offer an explanation for across-region variation, they leave the question of why, in the same region, practice variation across physicians can exist, i.e. why doctors that are performing in the same region differ significantly (Epstein & Nicholson, 2009).

Besides patients' preferences and regional factors, two of the most common explanations that are offered for sub-optimal medical practices are physicians' financial incentives and/or their risk avoiding behavior. Financial incentives have been argued to be a reason for over–utilizing resources. For example fee–for–service models are argued to create more incentives to perform more services in contrast to other financial systems (Bodenheimer & Grumbach, 2005). It is expected that the financial systems that give more incentives to physicians to overprescribe medical tests and treatments lead to inefficiencies, and the difference in financial systems and physicians' personal financial interests cause practice variation.

Non-financial reasons are also offered to explain the sources of sub-optimal behaviors. One of the common explanations for practice variation is the physician-specific factors such as risk aversion. These factors are known to persist over time, and have been argued to be difficult to measure outside of the laboratory (Epstein & Nicholson, 2009). For example, one way to avoid any risk of making a wrong decision is to prescribe medical tests for a larger population of patients. In such cases, physicians who are more risk averse would prescribe more tests. Furthermore, uncertainties have been argued to contribute to imperfect decisions and disagreements across doctors. Uncertainties can make it difficult to come up with a consistent decision, and it can result in more risk aversion. For example, in a high uncertain situation, higher risk aversion can lead to more defensive practices through abundant prescription of medical tests.

The Current Study

Physicians' decision models can be affected by observing the result of their past decisions. In psychology, such a learning process is referred as experiential learning and exists in many natural settings (Cyert & March, 1963; Levitt & March, 1988; Nelson & Winter, 1982). Although outcome feedback in general can help people to learn, there is a lot of other evidence that unclear, asymmetric, and delayed feedback can lead to sub-optimal decisions rather than the best possible decisions (Denrell & March, 2001; Elwin, Juslin, Olsson, & Enkvist, 2007; Fischer & Budescu, 2005; Ghaffarzadegan & Stewart, 2011a, 2011b; Huber, 1991; Lant, 1992; Levinthal & March, 1993; Miner & Mezias, 1996; Rahmandad, 2008; Rahmandad, Repenning, & Sterman, 2009; Stewart, Mumpower, & Holzworth, 2011).

This study offers a new explanation for practice variation beyond the current arguments. Controlling for most of the already discussed factors in the literature, we argue that the suboptimal medical practices in the form of heterogeneity and bias can appear through daily practices due to the characteristics of medical tasks. We hypothesize that outcome feedback, the process of judging effectiveness of different styles of practice, and the process of experience accumulation when combined with environmental uncertainties lead to heterogeneity in medical practice and bias toward overutilization of defensive practices. Through a simulation experiment, we show that for (mathematically) similar physicians visiting a similar population of patients, bias and variation emerge as physicians practice, receive information, and gain experience. In such cases, it is not necessary to assume different doctors have different financial incentives and preferences, or different personality traits. Instead, our findings indicate that sub-optimal decisions in the form of bias and variation can emerge as a result of task characteristics and daily practices. A preliminary validation of the results is achieved by comparing them with the data from cesarean section surgery in the states of New York and Florida. In the next sections, we review the case, the model, and simulation runs.

2. SUB–OPTIMALITY ACORSS OBSTETRICIANS

There are several reasons for obstetrics to be an important case of sub-optimal decisions in medicine. Obstetrics is prone to suboptimal decisions in the forms of bias and variation in practice. Cesarean section surgery has been argued to be over-performed (in 2007, 31.8% of birth cases in US (Hamilton, Martin, & Ventura, 2009)), and in many cases for nonmedical reasons (O'Callaghan, 2010). Patients get admitted to c-section surgery either through a pre-scheduled process where their doctor suggests the surgery before the due date, or during the vaginal delivery when the doctor decides to switch to a c-section surgery as a result of a new diagnosis. In additions to the costs of surgery and longer stays in hospitals, cesarean section surgery can cause more risks for healthy mothers and babies (O'Callaghan, 2010). While it is difficult to find optimal c-section rate, the 31.8% rate in US is much higher than the rate suggested by the World Health Organizations for the developed countries, i.e., 10-15% (World Health Organization, 1985).

In addition, it is found that physicians differ in their tendency to schedule a cesarean section surgery. The observation has been robust in many studies that control for patients' health status with different indicators showing that variation is more a practice style issue. Epstein and Nicholson (2009) investigate variation in cesarean surgery rate in New York and Florida. Controlling for patients' risk factors, they show that within and across regions there is a considerable variation in caesarean section surgeries and some physicians are more inclined to conduct surgery than others. They show that variation within a region is two times more than variation across regions. They estimate the standard deviation of the distribution to be 6.5 percentage points. They also find that in 24 percent of the cases, the physicians' risk adjusted c-section rate is statistically different from the regional mean at the five percent level. This is in contrast to many other studies that claim regional differences are the main causes of variation in practice. Interestingly, the variation in the style of practice persists over time and they claim that physicians do not converge to a community standard.

An obstetrician may perform more than 100 surgeries a year and one may expect that through these practices, she should learn about optimal decisions and make more accurate judgments about her patients. However, there are many complexities in the obstetric practices which make it difficult to learn.

One of the major sources of complexities is about the way that an obstetrician observes her decision outcome and the way the outcome is interpreted. In fact, in the context of obstetrics outcome feedback is asymmetric and contingent upon the decision, what is usually referred as conditional feedback in the literature of behavioral decision making. Let's assume that patients can be categorized into two groups, the ones with higher health risks that should go under c-section surgeries and the ones with lower risks that can deliver through vaginal birth. A doctor is not necessarily able to differentiate patients based on their true status, but she makes a judgment based on available data and she might make a wrong decision.

Based on an obstetrician's decision, which can be to conduct a vaginal birth or a c-section surgery, four decision outcomes as shown in Table 1 can happen which are true positive, false positive, true negative, and false negative. In addition, as shown in the table, outcome feedback in these four conditions is not similar. In the case of vaginal birth when the decision is a false negative, a doctor can observe her decision result during the practice and may even change her decision and conduct a c-section surgery for a patient that is in labor. But under false positive decisions, when a surgery is scheduled, c-section surgery will be conducted any way, and there is little clear outcome feedback. In sum, an immediate and clearer feedback exists on vaginal birth, but in c-section surgery, a high portion of poor outcomes will not be observed and may be attributed to patients' health risks.

		Obstetrician's initial decision	
		Vaginal birth (*clearer feedback*)	C-section surgery (*less feedback*)
True status of patient	Patients that should have c-section (High risk patients)	False negative In most cases, the obstetrician will observe the decision outcome immediately and sometimes may change the decision to a c-section surgery	True positive A correct decision. If the decision is performed well, it should have a proper outcome.
	Patients that should have Vaginal birth (Low risk patients)	True negative A correct decision. If the decision is performed well, it should have a proper outcome.	False positive An unnecessary surgery with possible side effects. However, feedback is unclear, delayed and can be attributed to other factors than a wrong decision.

Table 1: Four possible outcomes for an obstetrician's decision about vaginal birth vs. c-section surgery for different patients

Table 1 represents four different decision outcomes and conditionality of outcome feedback in the context of obstetrics. Feedback asymmetries exist in many other medical contexts. In the following, we focus on the cesarean section surgery as our case for modeling to make more sense of different concepts and variables in the model. We will later discuss how a general theory of practice variation can be developed based on the lessons from this modeling practice.

4. MODELLING

4.1. Conceptual model

Our unit of analysis is a physician.[1] An overview of the conceptual model is depicted in Figure 1.

[1] Although in many cases aggregation is encouraged, here, we are interested in the "distribution" of physicians to study practice variation, and prefer a disaggregated model of each physician. But the model belongs to the system dynamics modeling thread as it incorporates feedback-loops. For more information on the proper level of aggregation, see (Ghaffarzadegan, Lyneis, & Richardson, 2011; Rahmandad & Sterman, 2008)

The basic logic of the model is as follows. In each time period, a doctor visits a patient with a level of health risk and obtains *information about the patient*. Then the doctor compares the information with her *decision threshold* to perform a c-section. If the perceived risk is higher than the threshold, the doctor makes and performs a c-section *decision*, otherwise a vaginal delivery *decision*. The practice results in *practice outcome*. The outcome depends on how good the doctor's decision was for this specific patient, and how well the surgery was performed. Then, through a specific form of outcome feedback that was explained in Table 1, the doctor interprets the result of her decision. Outcome *feedback is more available* on vaginal delivery than c-section. Once the result is interpreted, if the perceived result is not good enough, the doctor tries to learn and correct her threshold. Through the whole process, the doctor accumulates *experience* of vaginal birth and c-section surgery, which in turn affects her skill of performing deliveries.

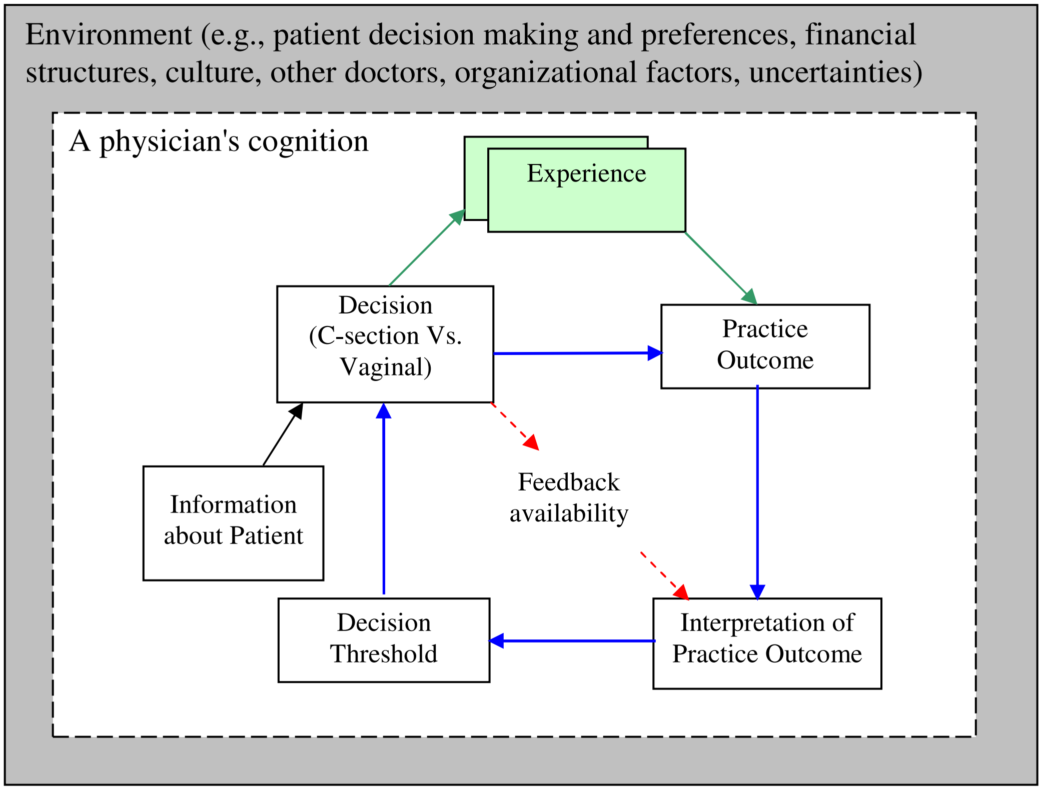

Environment (e.g., patient decision making and preferences, financial structures, culture, other doctors, organizational factors, uncertainties)
A physician's cognition
Experience
Decision (C-section Vs. Vaginal)
Practice Outcome
Feedback availability
Information about Patient
Decision Threshold
Interpretation of Practice Outcome

Figure 1: A simplified causal diagram of the medical practice model for obstetrics practice for a single obstetrician

4.2. Model Formulation

The depicted conceptual model in Figure 1 is formulated in three steps: decision making process (in the conceptual model, from *decision threshold* to *decision*), decision payoff (from *decision* to *experience* and *practice outcome*), and learning process (from *practice outcome* to *decision threshold*).

4.2.1. Decision Making Process

Decisions to perform a surgery are assumed to be based on information obtained from a patient. A physician decides to perform a c-section surgery if she diagnoses the level of risk to be high enough. Equation 1 represents this decision making principle. For each doctor, the model considers a threshold for performing c-section, and if the doctor perceives her patient's health risk to be higher than the threshold, she makes a c-section decision. In Equation 1, decision = 1 represents a c-section decision and zero represents a vaginal delivery decision.

$$\begin{array}{ll} \text{Decision} = 0 & \text{if Observed Health Risk} < \text{Threshold} \\ \text{Decision} = 1 & \text{if Observed Health Risk} \geq \text{Threshold} \end{array}$$

Equation 1

Observed Health Risk is a physician's observation of a patient's health risk. Equation 2 represents the variable. We assume a normally distributed random error in diagnosis, ε , with the mean of 0, that is no systematic bias in observation, and the standard deviation of ErrStdev.

$$\text{Observed Health Risk} = \text{Health Risk} + \varepsilon$$

Equation 2

Health risk represents patients' health risk normally distributed between 0 and 1, with the mean of HRMean and the standard deviation of HRStdev. Health risk =1 represents the worst health risk condition.

4.2.2. Decision payoff

As stated, in our model there are two alternatives for a doctor on each case: to perform a c-section or to perform a vaginal delivery. The outcome of the practice can depend on many factors, two of the most important ones are a) the decision match (i.e., how a decision for a patient matches that specific patient's health risk), and b) how well the decision was performed by the physician (i.e., a physician's skill in performing the decision).

$$\text{Practice Outcome} = \text{Normal Outcome} \times \text{Effect of Doctor's Experience} \times \text{Effect of Decision Match}$$

Equation 3

For the effect of Experience we use the learning curve idea and formulate it as shown in Equation 4. We consider two independent types of experience: experience of vaginal delivery (Experience_0) and experience of c-section (Experience_1).

$$\text{Effect of Doctor's Experience} = \left(\frac{\text{Experience}_i}{\text{Normal Exp}}\right)^{\alpha} \qquad 0 \leq \alpha \leq 1.$$

Equation 4

where *i* is equal to 0 for vaginal delivery and is equal to 1 for c-section. α is the sensitivity of practice outcome to experience and is between 0 (no effect from experience on practice outcome) and 1. The effect of decision match is a function of both decision and the level of health as shown in Equation 5,

$$\text{Effect of Decision Match} = f(\text{Decision}, \text{health risk})$$

Equation 5

where f should be defined in a way that f(0, y) represents how proper a vaginal delivery decision is for a patient with the health risk of y, and f(1, y) represents how proper a c-section surgery is for a patient with the health risk of y. We expect that a vaginal delivery decision to be more appropriate for the lower levels of health risk, and a c-section surgery to be more appropriate for the higher level of health risk. In other words, for Decision=0 we have $\frac{\partial f}{\partial y} < 0$, and for Decision =1 we have $\frac{\partial f}{\partial y} > 0$. We assume $f(x, y) = (1 - |x - y|)^{\beta}$, where β is a parameter representing the sensitivity of practice outcome to decision match and is between 0 (no effect from decision match) and 1.

Finally, physicians gain experience as they practice. We assume they accumulate experience as they perform the relevant decision, and they forget with the rate of (1/Time to forget). Equations 6a and 6b represent experience and skill in our model.

$$\Delta \text{Experience}_1 = \text{decision} - \frac{\text{Experience}_1}{\text{Time to forget}}$$ Equation 6a

$$\Delta \text{Experience}_0 = (1 \text{ - decision}) - \frac{\text{Experience}_0}{\text{Time to forget}}$$ Equation 6b

4.2.3. Learning Process

We assume, when feedback is available, physicians respond to the results and try to correct their decision threshold. In our model, we assume under vaginal delivery, feedback is provided *p*(Feedback Availability=1)=1, but under c-section, only very few times feedback is provided. Mathematically, if a c-section surgery is performed we have *p*(Feedback Availability =1)=q < 1 , that is in q portion of c-section surgeries feedback is provided.

When a physician receives feedback, she compares the result with her desired practice outcome. Higher deviation from the desired outcome results in higher threshold adjustment force (Adj Force). The change in threshold, therefore, will be in the direction of the observed risk: if she sees that her patient with a health risk lower than (resp. higher than) her threshold did not perform well under vaginal delivery (resp. c-section), the physician understands that her threshold was too high (resp. low), and for next patients she should decrease (resp. increase) her decision making threshold. Threshold adjustment force (Adj Force) also depends on Normal Adjustment Force (Normal Adj Force), representing one's normal speed of changing threshold.

$$\text{Adj Force} = \text{Noraml Adj Force} \cdot \text{Max (Desired Practice Outcome} - \text{Practice Outcome, 0)}$$ Equation 7

$$\Delta\text{Threshold} = (\text{Observed Health Risk} - \text{Threshold}) \cdot \text{Adj Force}$$ Equation 8

Desired Practice Outcome can be set in different ways. We assume Desired Practice Outcome to represent the maximum of one's average practice outcome and some acceptable norm (Minimum Acceptable Outcome).

$$\text{Desired Practice Outcome} = \text{Max}(\text{Average Outcome}, \text{Minimum Acceptable Outcome}) \qquad \text{Equation 9}$$

Average Outcome is the Moving Average of last two months practices.

5. SIMULATIONS

5.1. Base Run

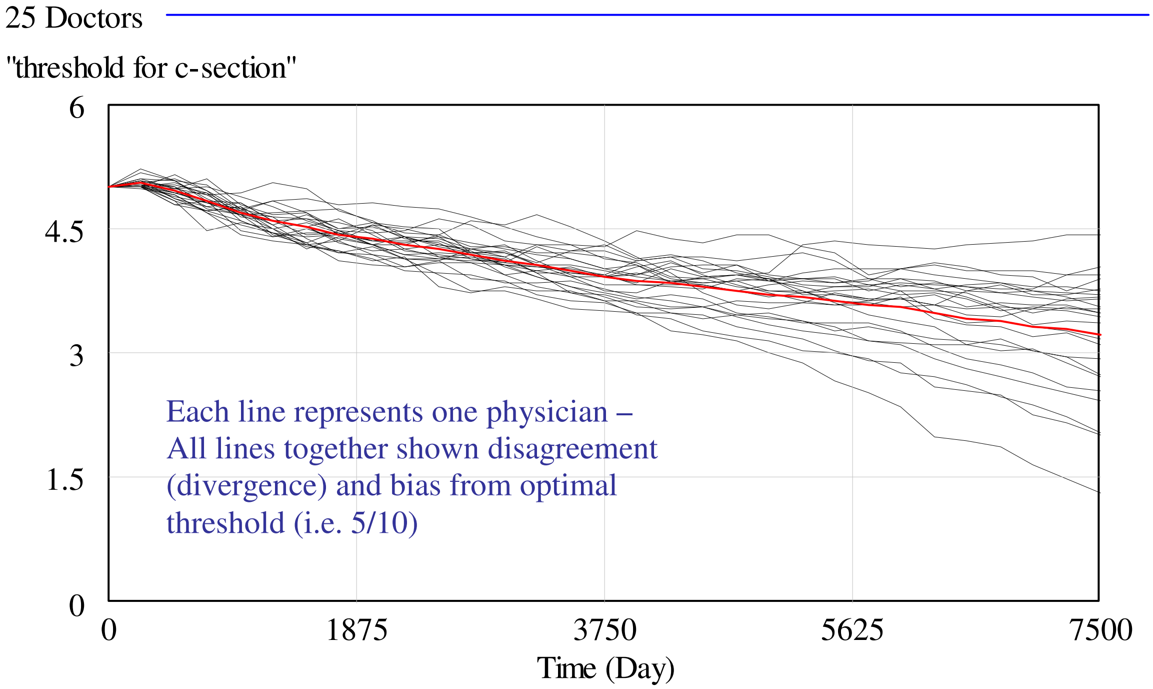
First, we present the base run of the model. Our time horizon is 30 years, each year 250 workdays, and our time unit is a workday, and we assume one baby delivery is performed per workday (about 5 baby deliveries per week, close to the real average number in the field). The model parameterization is reported in Table 2. In short, the parameters are set in a scale that the optimal threshold to conduct c-section is on health risk of 0.5, and the distribution of health risk among patients is in a way that the optimal c-section rate is 20%. Scaling the parameters is hypothetical, qualitatively set to represent a case close to the distribution of c-section surgery in the states of New York and Florida in section 2. We will discuss the sensitivity of the results to the parameters.

Parameters	**Values**
α, sensitivity of practice outcome to experience	0.5
β, sensitivity of practice outcome to decision match	0.25
Time to forget	50 workday
Normal Adj Force	0.01 1/workday
Minimum Acceptable Outcome	0.5
Normal Exp	1
Normal Outcome	1
q, portion of c-section surgeries where feedback is available	0.05
HRMean, mean of patients' health risk	0.3
HRStdev, standard deviation of patients' health risk	0.2
ErrStdev, standard deviation of error in a physicians' observation	0.1

Table 2: Parameters and functions for the base case simulation

Figure (2) shows the results for 25 randomly selected doctors, each graph representing one doctor. The x-axis is the time horizon, i.e., days of practice, and the y-axis is the threshold to make a c-section decision (how much health risk should be perceived to perform a c-section). Lower threshold would mean higher percentage of c-section deliveries, of which some would not be justified based on true risk. Therefore, each graph on Figure (2) shows how the threshold of a doctor changes during her practice. As we mentioned the model is parameterized in a way that 5/10 is the optimal threshold (a patient whose health risk is below 0.5 is in general better off with vaginal birth, vice versa). All physicians are assumed to start from an identical condition with the initial threshold of 5/10.

As we see the thresholds of the doctors go below the optimal threshold meaning that they perform c-sections more frequently than what is set in the model to be optimal. Further, as we see in the Figure, threshold to perform a c-section is different across different physicians, and they diverge as they practice. In other words there is a variation of c-section threshold.



Each line represents one physician – All lines together shown disagreement (divergence) and bias from optimal threshold (i.e. 5/10)

Figure (2): Threshold dynamics in the base run for 25 random doctors over 30 years of practice (250 workdays per year)

We can also look at the cross sectional synthetic data, output of the simulation model. In the real world, not all doctors start practicing together. For example, in the year 2011, some doctors are at their first year of practice, some at the second, and so on. Figure (3) shows the distribution of c-section surgery for 30,000 doctors, and compares that with the optimal threshold.

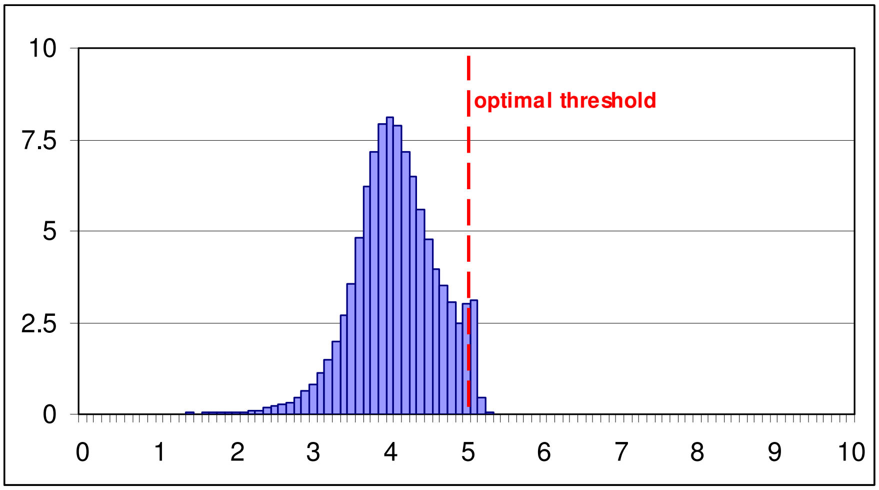


Figure (3): distribution of practice threshold and the optimal threshold.

As we see in Figure (3), the average of the distribution of threshold is smaller than the optimal threshold, showing that there is a bias in c-section rate toward performing more c-sections. In addition, the figure shows that there is a considerable variation across physicians. While there are physicians who set their thresholds below 3 (they perform c-section for patients with the health risk of 3/10), some set their thresholds around 5. It is important to mention that all of these physicians started from a similar initial condition including medial trainings and visited the same population of patients, and what was different across them was just a random seed that was influencing their errors and the order of the patients they visited. The distribution has also a relatively small peak around 5, as we start the initial condition from threshold = 5/10 and we sample from the whole population of the doctors including inexperienced doctors with the initial threshold of 5/10.

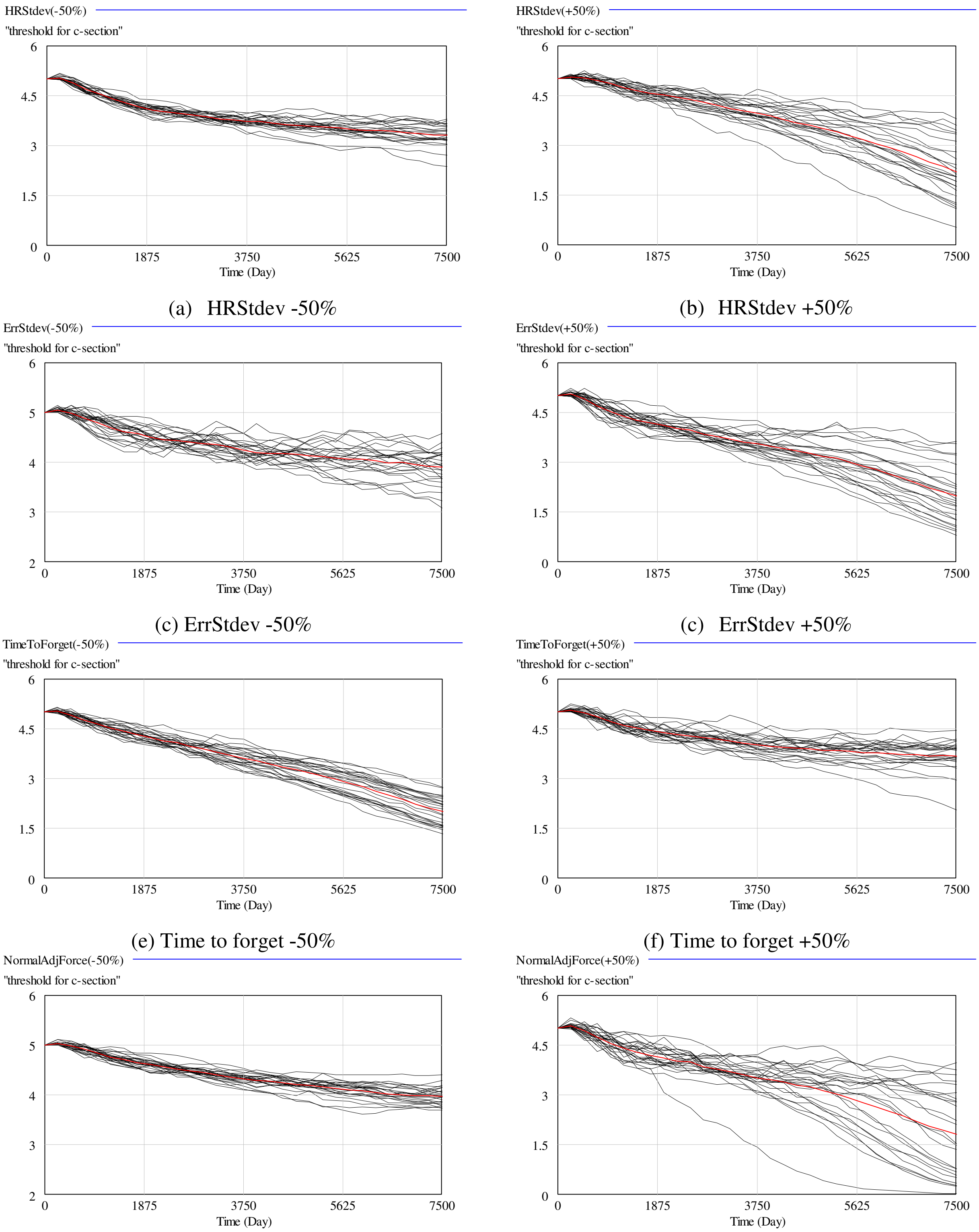
Next we report the results from simulating the model for a wider range of parameters. Figures 4a to 4f show the results. We test the base run simulation by changing the magnitude of four main parameters. First in figures 4a and 4b we change the variation in the society health risk (HRSdev) by ±50%. As we see increasing the variation in health risk results in more variation and more bias in c-section threshold across doctors. This result is not intuitive as all doctors are seeing the same population of patients, but in different order. Therefore, we see variation in patients' health results in variation in practice, even if physicians visit the same population of patients with the same average of health risk.

Figures 4c and 4d show the results for ±50% change in the standard deviation of ε in equation 2 (ErrStdev). The changes represent different levels of accuracy in physicians' observation of information about their patients (for ErrStdev=0, doctors are accurately observing patients' health risk). As we see in these figures, more errors result in a larger variation in practice and a larger bias. In both conditions, we still observe practice variation and bias.

In figures 4e and 4f, we change *Time to forget* by ±50%. As we see, a higher *Time to forget* results in less bias, but more variation. In fact, increasing *Time to forget* increases the steady state value of experience, and increases the effect of skill on practice performance, making the model more sensitive to experience in contrast to decision match. That means performance of a decision is relatively more important than a good match between the decision and the patient. In both conditions, still we see both practice variation and bias.

Finally, Figures 4g and 4h test changes in Normal Adjustment Force (NormalAdjForce) by ±50%. A lower Normal Adjustment Force represents a slower rate of change in threshold, and therefore, as we expect less bias and less variation at the end. As we expect, a quicker respond to feedback results in more changes in thresholds, and more bias and variation in practice.

These figures, overall, demonstrate that the base run results are not qualitatively sensitive to the parameters, and for a wide range of values for parameters we will still get both bias and variation in medical practice as results of experiential learning. However, the magnitude of bias and variation can be different under different scenarios. Therefore, the model demonstrates that bias and variation can arise solely from experiential learning and in the absence of financial incentives or personality traits. Next we compare the results of the model with the data reported in Epstein and Nicholson (2009).



(a) HRStdev -50%

(b) HRStdev +50%

(c) ErrStdev -50%

(c) ErrStdev +50%

(e) Time to forget -50%

(f) Time to forget +50%

(g) Normal Adj Force -50%

(h) Normal Adj Force +50%

Figure 4: Testing the base run simulation for a range of change in parameters: HRSdev±50%, ErrStdev ±50%, TimeToForget±50%, and NormalAdjForce ±50%.

5.2. A Comparison of the Model Results with the Data

As discussed in section 2, Epstein and Nicholson (2009) find a considerable level of practice variation in the states of New York and Florida. A comparison of our synthetic data from simulation with their data, although a weak test, can be interesting. We calculate the c-section rate for each data point and find the deviation from the mean for each data point, in a same way that is done in Epstein and Nicholson (2009). The results are shown in Figure 5. The resulted distribution is similar to their result (Figure 3a, in Epstein and Nicholson (2009, p. 1134)), and has a 10.5 percentage point standard deviation fairly higher than Epstein and Nicholson's observation (6.5 percentage point).

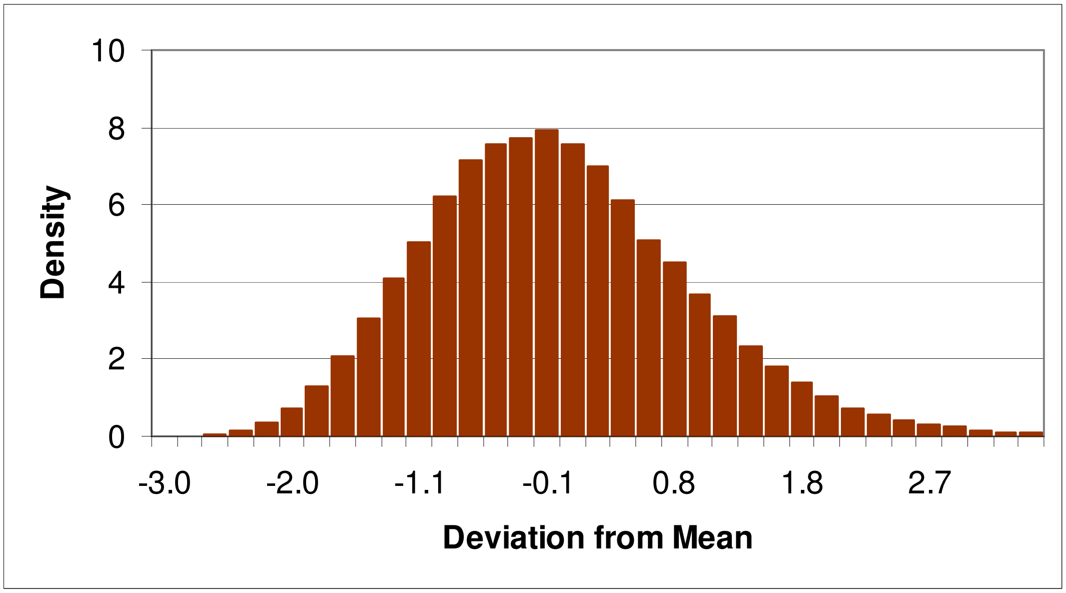


Figure (5): Distribution of deviation of c-section rate from mean as an effort to replicate Epstein and Nicholson's (2009) empirical finding. The standard deviation is 10.5 percentage points.

We believe although this practice is a weak test, but still the results are promising, and further detailed investigations can prove useful. There are a few parameters that need to be calibrated, and in such a case we can get a better fit. For example, the physicians' confidence on their negative decisions was unknown for us, and we used estimations from Elwin et al. (2007). Access to the detailed data on practice performance can be helpful.

5.3. Behavioral Analysis

We would like to further analyze the results in Figure (2) and find what parts of the structure result in practice variation and what parts are contributing to bias. We divide the model into two major sub-structures. The first sub-structure, which we call *the skill sub-structure* consists the model excluding the conditionality of feedback (providing full feedback after any decision whether it is vaginal or c-section surgery, i.e. q=1). The second sub-structure, which we call *the* *conditional feedback sub-structure*, is the entire model excluding the effects of skill (α=0). These experiments can give us a clearer explanation about “why” we have the results shown in Figures (2-5) (bias and variation), and what are the effects of each sub-structure on the final behavior.

Effects of Skill

First, we focus on the effects of skill, and assume a full feedback condition. To conduct this experiment, we provide feedback to physicians, after any kind of practice, whether it is a c-section or a vaginal birth.[2] Figure (6) shows different simulation runs for different random seeds. Each line represents one doctor’s decision threshold for performing c-section. The deviation shows that physicians are diverging as they gain more experience and some are more likely to perform c-section surgery than others. The graph shows that the stated rules are adequate to generate disagreements across doctors, even if they have similar financial incentives and similar initial training and even if clear feedback is provided after practice.

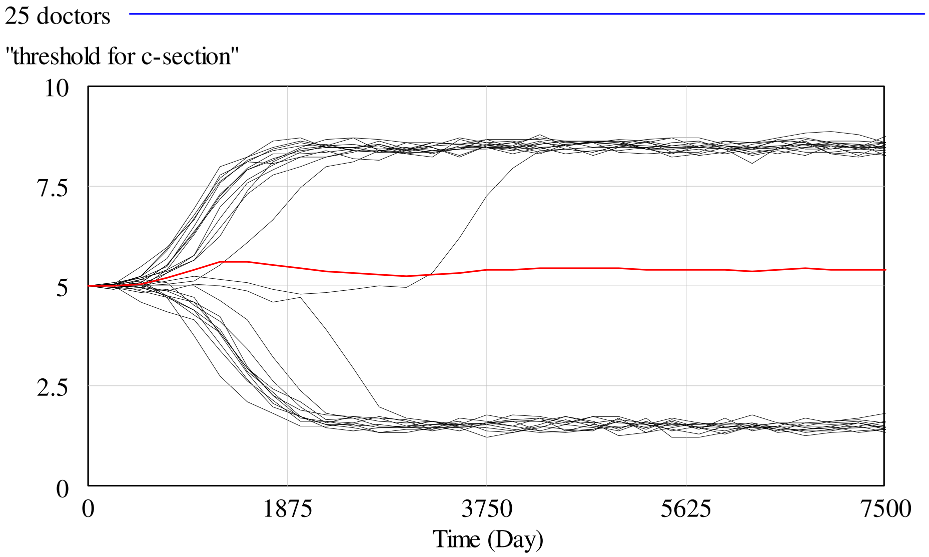


Figure (6): Dynamics of decision threshold for performing c-section for different doctors when feedback is complete, but performance depends on experience and skill.

The simulated doctors’ disagreement on what is the best decision for a single patient is due to the effect of past decisions on experience, and the causal connections between experience, perception and next decisions. As doctors perform their practice, they gain more experience, however in an unbalanced way where some have more experience in c-section and some have more experience in vaginal delivery. The unbalanced experience affects their performance, their perception of how effective each style can be, and their next decisions.

[2] All parameters are the same as base run, but q=1 (full feedback), and HRMean=0.5. The latter change provides equal rate of experience accumulation when threshold is equal to 5/10.

It is also important to mention that the skill substructure does not create bias in practice. As we see in Figure 6 the average of the distribution (the thicker line close to threshold = 5) is roughly around the optimal threshold.

Effects of Conditionality of Feedback

Now we focus on the second substructure and investigate the effect of conditionality in feedback on the final results.[3] As it is shown in Figure 7, when skill has no effect on the physicians' performance, their threshold do not diverge but decreases as they perform more practices, creating bias toward more c-section surgeries (i.e., lower threshold for c-section surgery).

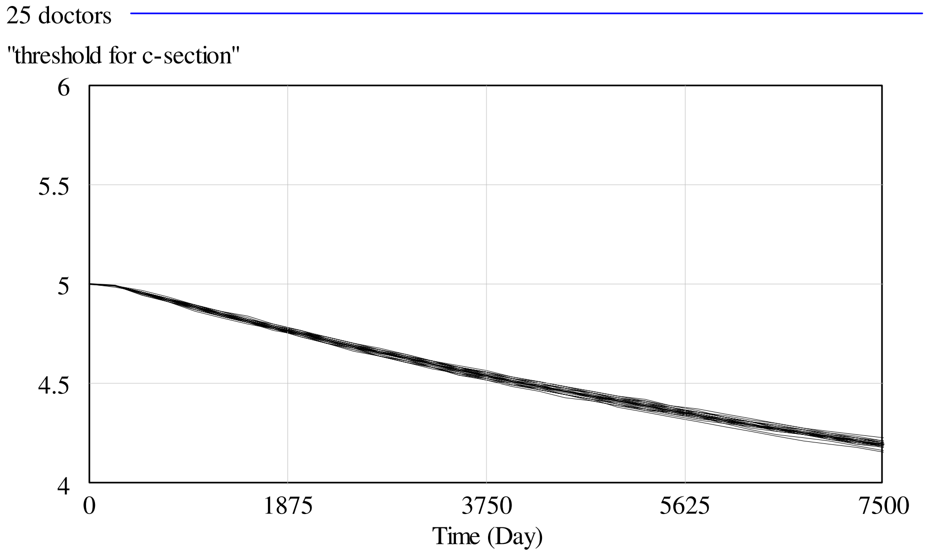


Figure (7): Dynamics of decision threshold for performing c-section for different doctors under the conditional feedback situation, controlled for the effects of experience and skill.

In summery, the analysis shows that the skill sub-structure results in practice variation and the conditional feedback sub-structure results in bias in practice. Next we examine the model for a more generic case.

5. 4. Simulation for a Generic Case

C-section is an illustrative example of bias and variation in medical decision making because like in many other domains, decisions are repetitive (frequent cases of patients), skill and experience are important on how medical decisions are performed, and there is an unbalanced outcome feedback. Many other medical domains have similar characteristics. However, in some domains there can be more (or less) elasticity to skill and some domains may provide more (or less) balanced outcome feedback. Having the mathematical model, we can generalize the arguments by changing the parameters and investigating the magnitude of bias and variation in different domains.

[3] All parameters are the same as base run, but α=0 (no effect from experience on decision outcome).

We define 6 (3×2) different hypothetical medical domains: 3 conditions on the sensitivity of practice outcome to experience (α= 0, 0.5, 1) times 2 feedback conditions (full feedback (FF), and conditional feedback (CF)). Although we have set these conditions hypothetically, they can represent different domains of medicine. For example, the condition of α=0 and CF can represent anti-biotic perception for otitis media in pediatric services. While a doctor should make a proper diagnosis and decision in prescribing anti-biotic, after the decision is made (the pills are prescribed) the doctor's skill has no effect on how anti-biotic pills affect the patient. In contrast, in a c-section surgery a doctor should make a good decision, but also should perform it well where her skill to perform a c-section surgery comes to play. Another example, can be in dental health and the practice of root channel, where a doctor's skill in performing a root channel is important (larger α), and feedback is conditional (usually a teeth with a removed nerve will not ache, even if the decision to remove the nerve was unnecessary). Practices that require frequent follow-ups and can control for the effects of decision can present a full feedback condition. For example a heart surgery can be considered as a case of a large α but a full feedback condition (FF), usually a physician will know if the surgery was performed well or not due to several next check-ups.

We simulated the model under the described 6 conditions (3 values of α times 2 feedback conditions). In each condition we have 1000 agents (physicians). We then calculate the magnitude of bias and variation in each of those conditions. Figure 8 compares the magnitude of bias and variation across those 6 conditions.

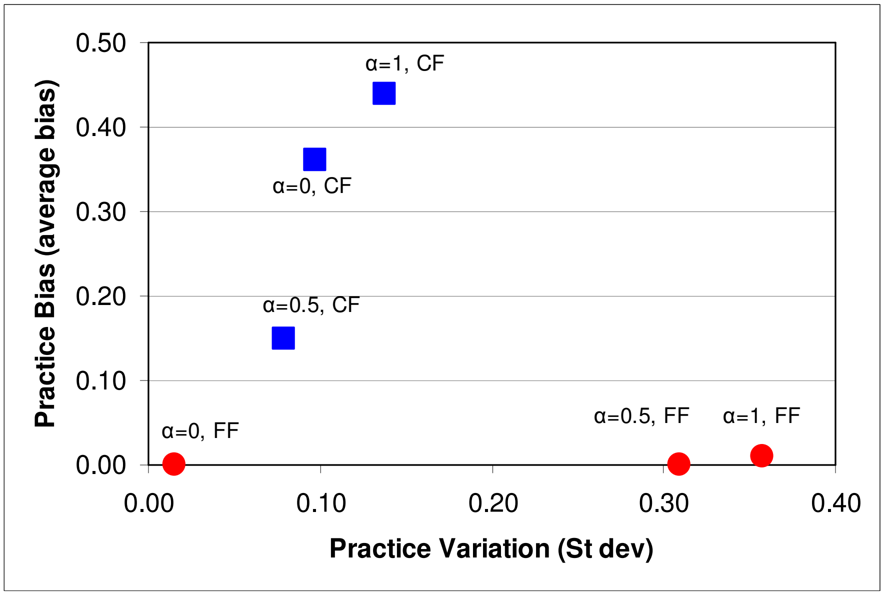


Figure (8): Bias and variation under 6 different conditions of medical practices: 3 different values for *sensitivity of practice outcome to experience* (α) times 2 feedback conditions (CF: conditional feedback, FF: full feedback).

Note: As we see, some conditions have more variation than others and some have more bias. For example the bias in practice is greater for (α=0.5, CF) than for (α=0.5, FF), while the latter has more practice variation.

Figure 8 shows that the magnitude of bias and variation under these conditions vary. In this figure, the x-axis is the magnitude of variation and the y-axis is magnitude of bias, and each point represents one of the 6 hypothetical domains of medicine. So, if a domain is closer to the origin of the graph (0,0), then there is less bias and variation there. For example the bias in practice is greater for (α=0.5, conditional feedback) than for (α=0.5, full feedback), while the latter shows more practice variation. The graph shows that under the full feedback conditions, larger elasticity to experience (α) results in larger practice variation. Interestingly, in a constant level of α, providing full feedback decreases bias but, in general, results in more variation (compare α=1 and CF with α=1 and FF). An empirical investigation of the findings can be interesting.

6. CONCLUSION

We develop a model of medical practice specifically tailored for obstetrics to study bias and variation in the practice of baby delivery when physicians should make a decision between a vaginal delivery and a c-section surgery for each of their patients. While the most common explanations for practice variation and bias are linked to patients' characteristics, regional characteristics, physicians' personality traits (e.g., the level of risk aversion) or their financial incentives, we offer a new theory of how medical task characteristics especially outcome feedback characteristics and the elasticity of the results to physicians' skills can result in practice variation and bias.

Our simulation model controlled for all of the common previous explanations offered in the literature and still created practice variation and bias. In our model, we do not impose any regional variation to our agents, and the personality characteristics and financial incentives of the agents are the same. In addition, doctors are visiting similar populations of patients with the same level of health risk. Therefore, the simulation results show that practice variation and bias does not have to be caused by patients' characteristics, regional characteristics, and physicians' personality traits and financial incentives, but can endogenously emerge through daily practices even across physicians with totally similar characteristics. In other words, the structure of medical tasks, and specifically physician's exposure to outcome feedback, and the experience accumulation processes through repetitive medical decisions can contribute to practice variation and bias.

We also analyzed the final results of the model through controlling different sub-structure of the model. The experiments revealed that accumulation of experience and skill can result in variation if one's performance is highly depended on one's skill. We also showed that the conditional feedback sub-structure drives bias in practice through forcing physicians to perform a kind of practice for which they receive less negative feedback. Further analyses revealed that the interactions of skill and conditionality of feedback exacerbate the bias, leading people to accumulate more experience on the kind of practice for which there is less negative feedback.

A preliminary validation of the results is achieved by comparing projected results with the actual data from cesarean section surgery in the states of New York and Florida. More empirical investigation is needed to test this dynamic theory.

The study contributes to the literatures of decision and policy sciences and medical decision making on different levels. First the study develops a new explanation for practice variation and bias in medicine. The new explanation is structurally different from the previous theories of practice variation and, therefore, has different policy implications. It is important to mention that this study does not reject any previous explanations offered for practice variation and bias in medicine, but it develops a new coherent theory of sub-optimal decisions in medicine. In other words, the theory is a new layer to existing understandings of the factors that contribute to bias and variation in medicine. This new theory needs to be empirically investigated. Second, our study is one of the first ones to apply the concepts of experiential learning into the studies of medical decision making. We believe such an approach is important as doctors perform repetitive tasks where for some portion of them they receive incomplete feedback. Third, the study has methodological contributions as it is the only study which has modeled physicians' decision making processes in a feedback-loop based approach. The model can be applied to a wide range of medical practices and can be used as a platform for further empirical investigations.

In short, we argue that practice variation and bias can dynamically emerge as physicians perform practices due to the learning characteristics of medical tasks. Such a structure can result in sub-optimal practices even if the financial incentives, the personality characteristics of the doctors, the regional characteristics, and the population of patients are totally similar.

ACKNOLEDGEMENTS

I am thankful to David Andersen for his helpful comments during this project. I am also thankful to Thomas Stewart, Erika Martin, George Richardson, Hazhir Rahmandad, Andrew Whitmore, Laura Black, Hyunjung Kim, and Luis Luna for their constructive comments. The model in the Vensim format is available through the online proceedings. All errors are exclusively the responsibility of the author.

References

Bodenheimer, T. S., & Grumbach, K. (2005). *Understanding Health Policy: A Clinical Approach* (4 ed.). New York: Lange Medical Books/McGraw-Hill.

Bynum, J., Song, Y., & Fisher, E. (2010). Variation in Prostate-Specific Antigen Screening in Men Aged 80 and Older in Fee-for-Service Medicare. *Journal of the American Geriatrics Society, 58*(4), 674-680.

Cyert, R. D., & March, J. G. (1963). *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall.

Denrell, J., & March, J. G. (2001). Adaptation as Information Restriction: The Hot Stove Effect. *Organization science : a journal of the Institute of Management Sciences, 12*(5), 16.

Elwin, E., Juslin, P., Olsson, H., & Enkvist, T. (2007). Constructivist Coding: Learning From Selective Feedback. *Psychological Science, 18*(2), 105-110.

Epstein, A. J., & Nicholson, S. (2009). The formation and evolution of physician treatment styles: An application to cesarean sections. *Journal of Health Economics, 28*(6), 1126-1140.

Fischer, I., & Budescu, D. V. (2005). When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks. *Organizational Behavior and Human Decision Processes, 98*(1), 39-53.

Fisher, E. S., Bynum, J. P., & Skinner, J. S. (2009). Slowing the Growth of Health Care Costs — Lessons from Regional Variation. *New England Journal of Medicine, 360*(9), 849-852.

Ghaffarzadegan, N., Lyneis, J., & Richardson, G. P. (2011). How small system dynamics models can help the public policy process. *System Dynamics Review, 27*(1), 22-44.

Ghaffarzadegan, N., & Stewart, T. R. (2011a). An Extension to the Constructivist Coding Hypothesis as a Learning Model for Selective Feedback When the Base Rate Is High. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(4), 1044-1047.

Ghaffarzadegan, N., & Stewart, T. R. (2011b). Stop, search, and (do not) learn: Barriers to learning in security contexts. University at Albany (SUNY).

Hamilton, B. E., Martin, J. A., & Ventura, S. J. (2009). *Births: Preliminary Data for 2007*. Hyattsville, MD: National Center for Health Statistics.

Huber, G. P. (1991). Organizational learning: The contributing processes and the literatures. . *Organization Science, 2*(1), 88-115.

Institute Of Medicine. (2000). *To Err Is Human: Building a Safer Health System*. Washington, DC: National Academies Press.

Institute Of Medicine. (2003). *Unequal Treatment: Confirming Racial and Ethnic Disparities in Healthcare.* . Washington, DC: National Academies Press.

Lant, T. K. (1992). Aspiration level updating: An empirical exploration. *Management Science, 38*(5), 623-644.

Levinthal, D. A., & March, J. G. (1993). A model of adaptive organizational search. *Economic Behavior and Organization, 2*(4).

Levitt, B., & March, J. G. (1988). Organizational learning. *Annual Review of Sociology, 14*, 319-340.

Miner, A. S., & Mezias, S. J. (1996). Ugly duckling no more: Pasts and futures of organizational learning research. *Organization Science, 7*(1), 88-99.

Nelson, R., & Winter, S. G. (1982). *An Evolutionary Theory of Economic Change*. Cambridge, MA: Harvard University Press.

O'Callaghan, T. (2010, Aug. 02, 2010). Too Many C-Sections: Docs Rethink Induced Labor. *Time*.

Rahmandad, H. (2008). Effect of delays on complexity of organizational learning. *Management Science, 54*(7), 1297-1312.

Rahmandad, H., Repenning, N. P., & Sterman, J. D. (2009). Effect of Feedback Delays on Learning. *System Dynamics Review, 25*(4), 309-338.

Rahmandad, H., & Sterman, J. D. (2008). Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Management Science, 54*(5), 998-1014.

Sorum, P. C., Stewart, T. R., Mullet, E., Gonzalez-Vallejo, C., Shim, J., Chasseigne, G., et al. (2002). Does choosing a treatment depend on making a diagnosis? US and French physicians' decision making about acute otitis media. *Med Decis Making, 22*(5), 394-402.

Stewart, T. R., Mumpower, J. L., & Holzworth, R. J. (2011). Learning to make selection and detection decisions: The roles of base rate and feedback. University at Albany (SUNY).

Sutherland, J. M., Fisher, E. S., & Skinner, J. S. (2009). Getting Past Denial — The High Cost of Health Care in the United States. *New England Journal of Medicine, 361*(13), 1227-1230.

Way, B. B., Allen, M. H., Mumpower, J. L., Stewart, T. R., & Banks, S. M. (1998). Interrater agreement among psychiatrist in psychiatric emergency assessments. *Am J Psychiatry, 155*(10), 1423-1428.

Welch, H. G. (2010). Screening Mammography — A Long Run for a Short Slide? *New England Journal of Medicine, 363*, 1276-1278.

Wennberg, J. E., Fisher, E. S., & Skinner, J. (2002). Geography and the Debate Over Medicare Reform. *Health Affairs, Web Exclusive*, W96-W114.

Wennberg, J. E., Freeman, J. L., & Culp, W. J. (1987). Are hospital services rationed in New Haven or over-utilized in Boston? *Lancet, 1*, 1185-1189.

Wennberg, J. E., & Gittelsohn, A. (1973). Small area variations in health care delivery. *Science, 182*, 1102–1108.

World Health Organization. (1985). Appropriate technology for birth. *2*, 436-437.