NBER WORKING PAPER SERIES

ON THE SORTING OF PHYSICIANS ACROSS MEDICAL OCCUPATIONS

Pascal Courty
Gerald R. Marschke

Working Paper 14502
http://www.nber.org/papers/w14502

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2008

We would like to thank Itai Agur, David Perez Castrillo, Javier Rivas and participants at ASSET07, EUI, University of Victoria, Berlin WZB, Leicester, Helsinki, and Collegio Carlo Alberto. All opinions and any errors are ours. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

On the Sorting of Physicians across Medical Occupations
Pascal Courty and Gerald R. Marschke
NBER Working Paper No. 14502
November 2008
JEL No. D82,I10,J31,J33,L23

ABSTRACT

We model the sorting of medical students across medical occupations and identify a mechanism that explains the possibility of differential productivity across occupations. The model combines moral hazard and matching of physicians and occupations with pre-matching investments. In equilibrium assortative matching takes place; more able physicians join occupations less exposed to moral hazard risk, face more powerful performance incentives, and are more productive. Under-consumption of health services relative to the first best allocation increases with occupational (moral hazard) risk. Occupations with risk above a given threshold are not viable. The model offers an explanation for the persistence of distortions in the mix of health care services offered the differential impact of malpractice risk across occupations, and the recent growth in medical specialization.

Pascal Courty
European University Institute
Via della Piazzuola 43, 50133
Firenze, Italy
pcourty@eui.eu

Gerald R. Marschke
John F. Kennedy School of Government
Harvard University
79 JFK Street
Cambridge, MA 02138
and NBER
jerry_marschke@harvard.edu

1 Introduction

The distribution of medical students' career choice among specialties is the object of intense debate. It is generally acknowledged, for example, that there is a chronic deficit of service in general medicine and in some specialties such as psychiatry (Brotherton et al. 2005).[1] If wages can freely adjust, as they often do in many countries, how can such imbalances persist? Another puzzle that has recently received much attention in the United States is the dramatic growth in specializations. The number of medical sub-specialties has grown from about 30 in the early 70's to more than 100 in the late 90's (Donini-Lenhoff, 2000). This paper develops a theoretical framework that explains how medical students sort across medical occupations, identifies a mechanism that can explain (a) the possibility of distortions in the supply of health care services across medical occupations and (b) changes in the extent of specialization, and establishes connections between these two outcomes.

Our approach assumes that medical occupations compete for talent but differ in their exposure to moral hazard risk. Thus, it is harder to evaluate effort in medical occupations where decision-making is less grounded in scientific fact and clinical evidence and where clinical outcomes are uncertain, difficult to compare, and rarely repeated. We develop this point in greater detail in the next section. Each occupation offers a pay-for-performance contract that trades-off compensation risk and effort incentive, as in the standard moral hazard model. When medical students select an occupation, they take into account how their performance will be evaluated and rewarded in their future career, and the associated risk they will face.

We investigate how differences in moral hazard risk across occupations influence the matching of occupations and talents, equilibrium contracts, and productivity. To our knowledge, this is the first work that considers the possibility that differential exposure to asymmetric information (moral hazard) across careers can influence career choices and the incentive to specialize. For the sake of concreteness and relevance, we study the

[1] Governments and medical organizations sometimes intervene with a variety of incentives and regulations (Thornton and Esposto, 2002).

market for graduate medical education, but one should keep in mind that many of our insights apply more broadly to other specialized labor markets.[2]

Specifically, the model combines moral hazard (Holmstrom and Milgrom, 1991), matching of physicians and medical occupations (Roth and Sotomayor, 1990), and pre-matching investments (Peters and Siow, 2002). We show that in equilibrium assortative matching takes place. More able physicians join medical occupations with low moral hazard risk, face more powerful performance incentives, and are more productive. Even when all medical occupations are identical ex-ante in terms of marginal productivity of effort, productivity is higher in less risky occupations. Two forces drive this result. First, more able physicians end up in occupations less exposed to moral hazard risk. Second, these occupations use more powerful incentives and this further magnifies differences in productivity. This second effect is best illustrated in the benchmark case where physicians are almost identical so that the first effect has a negligible impact on productivity.

The model identifies two channels through which an inefficient allocation of physicians can develop. To begin with, those occupations where risk is too high fail to emerge in equilibrium because they cannot produce enough surplus to cover outside options. Secondly, those physicians who accept a position in a high-risk occupation face less powerful incentives and supply less effort. These two channels imply that inefficiencies increase with occupational risk. In addition, this inefficiency differential across occupations increases as the distributions of physician talent and occupational risk are more dispersed. Since wages can perfectly adjust in the model, there is no shortage in the sense that some patients cannot find a physician. The distribution of consumption across occupations, however, is distorted relative to the first best one. There is under-consumption of high risk services.

In practice, many factors influence the sorting of physicians across medical occupations. For example, lifestyle and work schedule have been shown to influence career choice

[2]Examples include markets for other professionals (e.g. business and law education also offer specializations), and internal labor markets in large organizations (e.g. the army offers specializations such as combat, engineering, and intelligence and banks offer careers in sales, trading, and corporate finance). These specialized labor markets share in common the features that workers make specific human capital investments and commit to a career, while employers compete for workers with career promises.

(Landon et al. 2003a). Some of these factors also explain why the relative demand for some occupations can change over time. As long as wages can adjust, however, the factors that have been identified in the literature do not give rise to inefficiencies and should not raise concerns among policy-makers. Alternatively, some specialties may artificially restrict entry (Leffler, 1978), but this cannot explain the excess residency positions in some specialties. Finally, sub-optimal matching institutions may lead to inefficiencies, but it is not clear why systematic differentials across specialties should develop (Roth, 2008). Our contribution is to show that asymmetric information can generate systematic distortions in the distribution of health care supply across specialties.

After presenting the main results, we discuss several implications. The analysis suggests an explanation for the reduced enrollment in generalist career relative to specialist ones. Assume that moral hazard risk has decreased in specialty careers relative to generalist ones. This is consistent with evidence presented in DeWitt et al (1998), and could be due, for example, to a differential increase in the role of scientific measurement in many specialties. The implication is that specialty careers would become more attractive.

In addition, we argue that the model can help explain the growth in specialization over the past decades. The analysis shows that low-risk sub-specialities have an incentive to branch out from their main field. By doing so, they can attract better physicians and increase productivity. The model also sheds some light on the impact of malpractice reform on the distribution of physicians across specialties and across states (Kessler et al, 2005). Finally, an increased emphasis on performance measurement or on financial incentives, due to pressure from consumer advocate groups, health insurers or policy-makers as has happened in recent years, is likely to exacerbate the relative shortage of talent across medical occupations. Our analysis suggests that such reform should be implemented across-the-board to internalize externalities across specialties, instead of specialty-by-specialty, as currently done.

The incentive literature has studied performance measurement at the firm level (see Prendergast 1999 for a review) and many studies have investigated empirically the canonical proposition that incentive power should decrease with performance risk (Prendergast, 2002). Little attention has been dedicated to the study of broader implications of moral

hazard at a more macro level, across occupations or within a specialized labor market, as we do in this paper. In particular, there is to our knowledge no work investigating the possibility that moral hazard may lead to a failure to organize an economic activity.[3] Our central assumption that heterogeneity in the exposure to moral hazard risk may influence matching, plays an important role in the empirical literature studying the relation between incentives and performance risk (Chiappori and Salanie, 2003). While the empirical literature has focused on single occupations and considered only matching on risk aversion on the worker side (Ackerberg and Botticini, 2002), we consider differentiated occupations and study matching over worker talent.

Our model borrows two important ideas from the literature on organizational design. Holmstrom and Milgrom (1991, 1994) have shown the importance of interactions between different inputs of production and incentive instruments within a firm. Likewise, our model makes extensive use of complementarity, not only within production units as in the past literature, but also across units through assortative matching as in Besley and Ghatak (2005). Technically, the model is similar to Serfes (2005, forthcoming) who embeds Holmstrom and Milgrom (1991) within a matching setup, but he does so to capture the possibility of endogenous matching on risk aversion as suggested by the evidence from Ackerberg and Botticini (2002). In contrast with our model which assumes heterogeneity in talent, heterogeneity in risk aversion is not sufficient in general to guarantee assortative matching. We focus on heterogeneity in talent because there is much evidence showing that talent, measured for example as performance in medical school, determines the choice of specialty (e.g. Kiker and Zeh, 1998).

The next section provides some background discussion on the market for physicians. Section 3 presents the model. Section 4 derives the main results on sorting, productivity, and pay incentives. Section 5 discusses some implications and Section 6 concludes.

[3]The early transaction cost literature has explored the role of performance measurement in the organization of production (Alchian and Demsetz, 1972) but the focus of this literature is on the role of information cost in explaining the existence of firms.

2 Medical Occupations, Career Choice, and Moral Hazard

About one-third of physicians are primary care doctors. There are specialties within primary care such as pediatrics. When patients' specific health needs require further treatment, primary care physicians send them to see a specialist. Specialist physicians differ from primary care ones in that they focus on treating a particular system or part of the body, such as neurologists who study the brain. In the United States, there are about 30 medical specialties and 100 subspecialities. Different organizations are involved in controlling quality through accreditation of programs, certification and disciplining of physicians (specialty boards), and licensure (government).

Career Choice

The issue of matching physicians' choice of medical career with medical need is often debated and even more so when shortages become salient (Thornton and Esposto, 2002). Enrolment across careers displays cycles in addition to long term trends (Dorsey et al. 2003). For example, there has been a steady decline in the ratio of generalists to specialists over the past decade. Both the government and medical societies intervene, through funding priorities, subsidized loan programs, educational reforms, and regulated work schedules to name just a few examples, to correct trends that could have a negative impact on the ability to provide, in the long-term, a balanced specialty mix of medical care. For example, in 1993 and 1994, the Physician Payment Review Commission advised Congress to implement a system of quantitative restrictions on positions.

There is a large literature studying the choice of medical specialty both in medicine (e.g. Weeks and Wallace, 2002) and economics (Nicholson, 2002). Medical students have discretion over the choice of medical specialization. For example, Bland and Isaacs (2002) report that more than 40 percent of students select a specialty during or after their third year of medical school, and between 40 and 60 percent of medical students change their minds at least once. A large body of research has shown that demographic characteristics influence career choice. In addition, there is also much evidence that economic incentives matter. Among other considerations, perceived future earnings, educational debt,

expected lifestyle (work schedule and predictability of hours), and malpractice risk, have been shown to influence the choice of specialty.

Under the assumption that compensation can adjust, the fact that physicians have preferences over specialties cannot explain the existence of differential inefficiencies across careers. To single out the driving force in our mechanism for inefficiencies, the model will assume that all medical occupations are identical in all respects except in their exposure to moral hazard, and that physicians select a medical occupation only on the basis of expected future utility.

Health Care, Moral Hazard and Pay for Performance

The model assumes that moral hazard prevails in medicine. This is consistent with the view that physicians' output is notoriously difficult to measure. For example, Arrow's (1963) seminal analysis of medical care recognized that 'uncertainty as to the quality of the product is perhaps more intense here than in any other important commodity.' That physicians are providing different levels of care is well documented (Committee on Quality Health Care in America, 2001). Perceptions of quality are formed by experience, physician and hospital ratings (such as report cards, Dranove et al., 2002; and U.S. News and World Report rankings), peer assessments of physicians (the advice of a family doctor), reputation, word-of-mouth, disciplining boards, and so on. These perceptions drive up or down the demand for particular physicians.[4]

The model also assumes that employers can reduce moral hazard by introducing performance incentives but they have to strike a balance between incentives and risk exposure. Again, there is much evidence consistent with this view. For example, Gaynor and Gertler (1999) show that compensation methods respond to the riskiness of the environment, as we assume in the model, in addition to the more basic issue that performance incentives have a substantial effect on physicians' effort. Many performance measures are used in practice, such as the number of patients treated, billed hours, clinical outcomes, report cards, adherence to clinical guidelines, patient surveys, peer evaluations, to name just a

[4]More recently, a number of private firms and public organizations (e.g. National Committee for Quality Assurance, HealthGrades) have started to compile information on individual physicians' performance and are making it available over the Internet. One would expect ratings to influence decisions by patients and managed care organizations, and therefore physician demand.

few examples. There are several ways through which physician compensation is linked to performance. For example, fee-for-service and capitation contracts play a significant role under managed care (Gold, 1999), revenue sharing is common in partnerships (Gaynor and Gertler, 1999), and academic hospitals use measures of clinical work and academic standing to determine pay (Abouleish et al. 2005).

Heterogeneity across Medical Occupations

The departing point of the model is that the exposure to moral hazard varies across medical occupations. To fix ideas, we discuss different sources of heterogeneity that are consistent with this view. There is much agreement that the ability to assess the quality of physicians' decision-making varies across treatment areas. For example, outsiders can observe whether a mammography or childhood immunization was prescribed for a woman or child who falls within the guidelines of recommended practice, or whether a β-blocker was administered to a patient after a myocardial infarction. In the treatment of complex illnesses, however, performance is more difficult to evaluate. For example, Angell and Kassirer (1996, p. 884) motivate this challenge by commenting that "treating congestive heart failure or urosepsis in a patient with diabetes mellitus, who may have other medical problems as well, involves not only a complex series of decisions and interactions, but also the nearly imponderable element of individual variation." In such cases, effort is likely impossible to evaluate on a case by case basis.

More generally, the information available on outcomes of care and clinical processes depends on the medical occupation.[5] Some clinical treatments have only a statistical impact while others have a deterministic one and the lag between action and effect varies greatly across treatments. For example, those specialists who repeatedly practice only a few, possibly highly complex, procedures are less subject to performance risk. This is known as the sample size effect and Landon et al (2003b, p. 1197) note that the "sample size problem is less difficult in certain specialties, like cardiac surgery and interventional cardi-

[5]Loeb (2004) reports that "not all decision-making in medicine is grounded in scientific fact and clinical evidence (i.e. opinion plays a significant role in medical decision-making). While evidence-based clinical practice guidelines exist in a variety of specialities and subspecialities in medicine, consistent evidence suggests that adherence to guidelines is poor."

ology, in which physicians may perform a large number of a limited range of procedures."[6] Taking into account this effect as well as many other considerations, they conclude that the ability to measure clinical performance varies across medical occupations.[7] Consistent with this view, a subcommittee hearing on measuring physician quality reports that "it does depend very much on the specialties. There is a very wide range of specialties and conditions for which administrative data–in particular when we include laboratory results and pharmacy–can provide a very solid picture of physician performance–not in all specialties."[8,9] Taken together, the evidence is consistent with our main assumption that there exist systematic differences across medical occupations in the quality of information available to assess physician performance, and therefore differences in occupational exposure to moral hazard.

3 Model

The objective of the model is to identify a mechanism that can generate differential inefficiencies across medical occupations and also to reveal the factors that cause this differential. To achieve this goal, we selectively include in the model the features that can

[6]To illustrate, consider the problem of evaluating a neurosurgeon's removal of a nerve root tumor, a tumor that develops from the cells of the nerve or of its lining near the point it exits the spinal canal. The outcome—the degree to which pain and numbness are reduced and whether permanent paralysis results—depends not only on the skill, dexterity, and care of the surgeon, but also on chance. It depends on chance because the surgeon has little influence over the size, location, and degree of separation between the tumor and nerve tissue but these things influence the surgical outcome. In specialized practices where this procedure is performed frequently, the noise averages out, and the quality of effort is more easily discerned. In more general practices (say in less urban areas) where the surgeon is treating a variety of conditions her performance will be noisier.

[7]"Few medical specialties have an evidence base that is robust and comprehensive enough to support physician clinical performance assessment. Some specialties such as cardiology and endocrinology have some evidence-based process measure that have been definitely linked to improved patient outcomes. Other specialties such as cardiac surgery, have outcomes that have been studied, such as mortality in coronary artery bypass grafting. Outcome measures for other specialties, however, occur too infrequently or tool long after care to make their collection feasible."

[8]Hearing on Measuring Physician Quality and Efficiency of Care for Medicare Beneficiaries. http://waysandmeans.house.gov/hearings.asp?formmode=detail&hearing=390

[9]For example, patient management plays an important role in medical care but the associated skills are very difficult to measure. One aspect considered in the medical literature corresponds to empathy. Many experts believe that empathy, defined as understanding the "patient's inner experiences and perspective and communicating this understanding", influences clinical outcomes (Hojat et al. 2002). The importance of empathy, however, varies across specialties, being more important in the "people-oriented" specialties (such as psychiatry, pediatrics or family practice) as compared to the technically-oriented disciplines (such as surgery or anesthesiology).

generate such an effect or magnify it. Because these features are not necessarily present simultaneously, the distortions observed in practice may not be as dramatic as those that the model can explain.

There are three building blocks to the model: pre-matching investments, matching, and moral hazard. The moral hazard part uses functional forms that are standard in the incentive literature and justified in Holmstrom and Milgrom (1991). Following the assortative matching literature, we model matching using unidimensional preference ordering keeping in mind that a more realistic model would acknowledge the fact that matching takes place along other dimensions.

The model has physicians and employers (hospitals, health maintenance organizations, partnerships). To simplify the exposition, we assume without loss of generality, that each employer represents a unique medical occupation, and as such, is endowed with a distinct technology to control moral hazard risk. Different medical occupations can offer different incentive contracts. This is consistent with the fact that the hospital or health organization where physicians work, can treat differently physicians in different occupations. There are three periods (see Figure 1). In period one, physicians invest in human capital and employers invest to reduce exposure to moral hazard risk. At the end of period one, the distribution of human capital amongst physicians and moral hazard risk amongst occupations are observed. In the second period, physicians and employers match and agree on a contract. In the third period, physicians exert effort, nature draws performance, and contracts are executed.

There is a continuum of physicians indexed by $\rho \in R = [\rho_0, \rho_1]$. Physician type ρ is distributed with density $f > 0$ and continuous distribution F. Investment in human capital lowers the cost of effort. All results follow if we assume instead that it increases productivity of effort and we will further discuss the issue after presenting the results. A physician with cost of effort c gets disutility $C(e|c) \geq 0$ for exerting effort $e \geq 0$ where $C_e > 0$, $C_{ee} > 0$, $C_c > 0$, and $C_{ce} > 0$. Physician of type ρ achieves cost index $c \geq 0$ if she invests $H(c|\rho) \geq 0$ where $H_c < 0$, $H_{cc} > 0$, $H_\rho < 0$, $H_{\rho c} > 0$, and $H = 0$ for c large enough. The utility of a physician of type ρ who selects cost of effort c, exerts efforts e,

and is paid wage w is

$$U^{3,P}(e, c, w|\rho) = -exp[-r(w - C(e|c) - H(c|\rho))]$$

where superscripts denote the period and agent considered, and r is the coefficient of absolute risk aversion. There is a continuum of employers (or medical occupations), indexed by γ, which are taken as given. γ is distributed according to density $g > 0$ and continuous distribution G with support $\Gamma = [\gamma_0, \gamma_1]$. Work effort is subject to moral hazard. Employers, however, can reduce exposure to moral hazard. Employer γ can achieve moral hazard risk $s \geq 0$ at cost $K(s|\gamma) \geq 0$ where $K_s < 0$, $K_{ss} > 0$, $K_\gamma < 0$, $K_{s\gamma} > 0$, and $K = 0$ for s large enough.[10] Each employer receives an imperfect measure of effort according to

$$m(e, s) = e + \varepsilon_s$$

where ε_s is an error term that is distributed normally with mean zero and variance s^2. The measurement errors are independently drawn across employers.

In period two, physicians and employers decide whether to match, and conditional on matching, agree on a contract. Following the literature, we restrict to linear compensation schedule $b = (b_0, b_1)$

$$w(m) = b_0 + b_1 m.$$

The physician then chooses effort level e and nature draws performance outcome m. Finally, the employer rewards the physician according to the agreed rule $w(m)$. We first assume that all employers equally value $\Pi(e) > 0$ effort level e such that $\Pi' > 0$ and $\Pi'' < 0$. We later discuss the case of heterogeneous productivity across occupations. Employers are risk neutral and maximize $\Pi(e) - Ew(m) - K(s|\gamma)$, or

$$U^{3,S}(e, s|\gamma) = \Pi(e) - b_1 e - b_0 - K(s|\gamma).$$

We focus on stable matching (Roth and Sotomayor, 1990). We denote $\mu^{2,S}(c)$ the employer matched with physician c if physician c is matched and $\mu^{2,S}(c) = \emptyset$ otherwise. For ease

[10] Investments to reduce moral hazard risk should be interpreted broadly. It could capture the investment made by the hospital hiring the physician. Alternatively, it could also be interpreted as medical societies investing in monitoring quality through re-licensure, and disciplining.

of exposition, we similarly define $\mu^{2,P}(s)$ and we have $\mu^{2,P}(\mu^{2,S}(c)) = c$ for matched pairs. The equilibrium contract function associates a contract $B(c) = (b_0(c), b_1(c))$ to each matched pair. The outside options of employers and physicians are $U^{0,S}$ and $U^{0,P}$ respectively. In stage two, we denote $u^{2,S}(s)$ the expected payoff of employer s and $u^{2,P}(c)$ the certainty equivalent continuation payoff of physician c.[11] Following Peters and Siow (2002), we define a rational expectation equilibrium as:

(1) A set of investment rules $c(\rho)$ and $s(\gamma)$ for physicians and employers that maximize their payoffs conditional on expectations about $u^{2,P}()$ and $u^{2,S}()$.

(2) The matching and contract functions $\mu^{2,S}(c)$ and $B(c)$ are stable. In period two, (a) no pair of physician and employer (c, s) such that $\mu^{2,S}(c) \neq s$ wants to match under any contract, (b) no pair of physician and employer (c, s) such that $\mu^{2,S}(c) = s$ wants to change contract.

(3) Period one participation says that no matched physician or employer prefers the outside option over the equilibrium payoff.

(4) An incentive compatible level of effort $e(c)$ for each matched physician.

(5) Physicians and employers have rational expectations: the functions $u^{2,P}()$ and $u^{2,S}()$ are consistent with $\mu^{2,S}(c)$, $B(c)$, and $e(c)$.

The functions $u^{2,P}()$ and $u^{2,S}()$ correspond to the market return of investments. As in Peters and Siow (2002), physicians and employers choose optimal investments given their expectations about the market returns. The main difference is that utility is transferable in our model so the functions $u^{2,P}()$ and $u^{2,S}()$ do not depend only on equilibrium matching, as would be the case under non-transferable utilities, but also depend on the equilibrium sharing rule.

Our objective is to derive equilibrium cross variations in $(c(\rho), s(\gamma), \mu^{2,S}(c), e(c), B(c))$. The main innovations of the model is to capture the fact that exposure to moral hazard varies across employers and to allow for matching. For the sake of generality, we have introduced the possibility of pre-matching investments by employers, and this addresses

[11] c is indifferent between receiving $u^{2,P}(c)$ for sure and matching with $\mu^{2,S}(c)$ under contract $B(c)$. As will become clear soon, CARA utility implies that the certainty equivalent does not depend on physician type ρ.

the concern that exposure to moral hazard is to some extent endogenous. We also consider pre-matching investment by physicians to capture the effort supplied during medical school training, but this feature of the model is not essential.

In addition to boundary conditions, two technical conditions are sufficient to demonstrate equilibrium uniqueness.[12]

Assumption 1: (A1a) $C_{ee}^2 + C_e C_{eee} > 0$, (A1b) $C_{ece} > C_{eee}$.

This assumption holds, for example, for quadratic cost $C(e|c) = \frac{ce^2}{2}$. Following Holmstrom and Milgrom (1991, p.179), we define $W^{2,SB}(c,s) = Max_e \left\{\Pi(e) - C(e|c) - \frac{r}{2}(sC_e(e|c))^2\right\}$ the period two information constrained surplus function of pair (c,s) in certainty equivalent monetary units.[13] A1b is sufficient to show that effort and moral hazard risk are complement in the joint surplus function.

Assumption 2: (A2a) $H_{cc} > W_{cc}$, $K_{ss} > W_{ss}^{2,SB}$. (A2b) $(H_{cc} - W_{cc})(K_{ss} - W_{ss}) > \left(W_{sc}^{2,SB}\right)^2$.

Assumptions A2 guaranty that the pre-matching investments are monotone in type.

4 Analysis

We derive the main qualitative results in the context of the general model. To discuss additional implications, we consider a restricted version of the model where it is possible to derive closed form solutions. We assume no pre-matching investments $\rho = c$ and $\gamma = s$ and functional forms $C(e|c) = \frac{ce^2}{2}$ and $\Pi(e) = \pi e$. The reader who perfers to start with an example should read section 4.3 first. All proofs are presented in the appendix.

4.1 Symmetric Information

As a benchmark, consider the case where effort is perfectly observable (no moral hazard). Employers do not invest to reduce s ($K = 0$) and sorting is arbitrary. Since employers are

[12] $C_e(0|c) = 0$, $C_e(\infty|c) = \infty$, $K_s(0,\gamma) = -\infty$, $K_s(\infty,\gamma) = 0$, $H_c(0,\gamma) = -\infty$, $H_c(\infty,\gamma) = 0$, and $C_c(0|c)$ bounded.

[13] This is the sum of the specialty profit and physician certainty equivalent. The meaning of this expression will become clear after Lemma 1.

identical, they receive the same payoff, which is determined such that both sides of the market are willing to participate, and physicians receive the residual surplus. Physician of type ρ chooses $c(\rho)$ and $e(\rho)$ such that, $H_c(c|\rho) + C_c(e|c) = 0$ and $C_e(e|c) = \Pi_e(e)$ independently of the employer she is matched with.

In the application with no pre-matching investments ($H = 0$) and quadratic cost of effort, the effort supplied by physician c is

$$e(c) = \frac{\pi}{c}.$$

The symmetric information period two surplus by pair (c, s) measured in monetary terms is $W^{2,FB}(c, s) = Max_e \{\Pi(e) - C(e|c)\}$. Using the specific functional form presented earlier, we have

$$W^{2,FB}(c, s) = \frac{\pi^2}{2c}.$$

Employers do not to invest to reduce moral hazard risk, s. Surplus increases in talent and is independent of employer risk. Medical occupations are completely undifferentiated. Under symmetric information, the model does not say anything about specialization. We will return to the issue when we discuss the incentives to specialize under asymmetric information.

4.2 Asymmetric Information

We analyze the problem backward. Consider a physician of type c who has matched in period two with employer s and agreed to contract (b_0, b_1). In period three, the physician sets e to maximize $b_1 e - C(e|c)$. The period three effort $e(c, b_1)$ solves

$$C_e(e|c) = b_1 \tag{1}$$

We can now characterize the incentive component of the period two contract.

Lemma 1: In stage two, any matched pair (c, s) agrees on incentive contract

$$b_1(c, s) = \frac{\Pi_e(e)}{1 + rs^2 C_{ee}(c|e)} \tag{2}$$

Lemma 1 says that any incentive contract that does not maximize the information constrained joint surplus of pair (c, s) cannot be part of an equilibrium. If this would be the case, pair (c, s) could renegotiate, agree on a contract with incentive parameter $b_1(c, s)$, and set a transfer payment $b_0(c, s)$ that makes both parties better off.

The role of CARA utility is now transparent. As in the standard principal agent model, CARA utility implies that the sharing rule $b_1(c, s)$ does not depend on the fixed transfer $b_0(c, s)$ and this makes the contract design problem separable in these two dimensions. In addition, CARA implies that the sharing rule is independent of the level of pre-matching investments $H(c|\rho)$. We get inter-period separability; we can solve for matching and contracting in stage two independently of the stage one pre-matching investment choices. The definition of $W^{2,SB}(c, s)$ becomes clear. In period two, physician c and occupation s agree on contract $b_1(c, s)$. $W^{2,SB}(c, s)$ corresponds to the maximum continuation payoff (in certainty equivalent units) that the pair can achieve under incentive compatibility.

We now turn to the matching problem. Stability requires $W^{2,SB}(c, s) = u^{2,P}(c) + u^{2,S}(s)$. For the sake of exposition, we initially assume that all physicians and occupations match. A sufficient condition for this to hold is $\ln\left(-U^{0,P}\right) = U^{0,S} = 0$. Matching is positive assortative (PAM) in period two if $\mu^{2,S}(c)$ is increasing and similarly in period one if higher γ match with higher ρ. We can now state our main result.

Lemma 2: In any equilibrium, there is PAM in (c, s) in period two and in (ρ, γ) in period one. Types (ρ, γ) such that $G(\gamma) = F(\rho)$ match together.

Two forces drive the PAM result. First, the physician cost of effort and occupational risk are complement in the joint surplus function $W^{2,SB}(c, s)$. This alone implies PAM in (c, s) in period two. Second, investments that lower the cost of effort and the level of risk are complement with types $H_{\rho c} > 0$ and $K_{\gamma s} > 0$. Combined with PAM in period two, this implies PAM also in period one. Clearly, the assumption of complementarity between investment and type characterizes the situation where pre-matching investments increases the amount of heterogeneity in (c, s) relative to the no-investment benchmark. Without complementarity, pre-matching investments may maintain or even reduce the

initial heterogeneity in (c, s). Still, for any distribution of (c, s) Lemma 2 shows that there is assortative matching in period two and this is what drives our main results on efficiency differential across occupations, as we will see soon. The main point is that the analysis is robust to pre-matching investments and the results are magnified under complementarity.

The outcome of sorting rests on the assumptions we made on the nature of heterogeneity amongst workers and occupations. Sorting is governed by the interaction between worker and occupation type in the joint surplus function. Workers may differ in other dimensions than ability and occupations may differ in other dimensions than risk. For example, Serfes (2005) assumes that workers differ in their degree of risk aversion r (he assumes that employers differ in riskiness s as we do in this paper) and shows that it is possible to characterize the equilibrium only in specific cases.[14] In contrast, we consider matching between worker ability and occupational risk. Since c and s are complement in $W^{2,SB}$ only PAM can occur.[15]

More generally, we could have assumed that worker ability is captured by their marginal productivity instead of marginal cost of effort. All results would follow if physicians would have identical cost function but would differ in term of marginal productivity (worker of type π produces $\Pi(e) = \pi e$). A central assumption is that worker ability is independent of occupational risk. The analysis may change if one assumes that part of the risk can be controlled by the worker. For example, the equilibrium matching may differ if more able workers can control risk more efficiently. The model, therefore, applies primarily to sources of risks that are outside the control of physicians. We can now state our main proposition.

Proposition 1: There exists a unique equilibrium up to the fixed constant $b_0(c(\rho_0))$.

[14]When risk plays a small (large) role in the sense that rs^2 is small (large) for the highest (lowest) types, then r and s are complement (substitute) in $W^{2,SB}$ and PAM (Negative AM) holds. He cannot characterize the equilibrium for intermediate ranges of rs^2.

[15]Our analysis sheds new light on the debate on the tenous link between risk and incentives (Prendergast, 2002). While the literature has exclusively focused on unobserved heterogenity in risk-aversion (Ackerberg and Botticini (2002), Serfes (2005)), our model shows that unobserved heterogeneity on worker ability also matters. Since ability is negatively associated with risk and positively with incentives, ignoring such heterogeneity introduces a bias toward over-estimating of the negative relationship between risk and incentives.

Period two matching is defined by

$$\mu^{2,S}(c(\rho)) = s(G^{-1}(F(\rho)) \tag{3}$$

contracting is defined by (2), efforts by (1), and investments by

$$\begin{cases} H_c(c(\rho),\rho) = W_c^{2,SB}(c(\rho), \mu^{2,S}(c(\rho))) \\ K_s(s(\gamma),\gamma) = W_s^{2,SB}(\mu^{2,P}(s(\gamma)), s(\gamma)) \end{cases}. \tag{4}$$

Equations (3) and (4) define the matching function and pre-matching investment functions. The sharing rule is defined by (2). The stability conditions define the period two continuation payoffs (up to a constant) according to $u_c^{2,P}(c) = W_c^{2,SB}(c, \mu^{2,S}(c))$ and $u_s^{2,S}(s) = W_s^{2,SB}(\mu^{2,P}(s), s)$ and the constant is determined by the allocation of surplus between the lowest pair $b_0(c(\rho_0))$ which can take any value so long as the participation constraints are satisfied. The resulting $u^{2,P}(c)$ and $u^{2,S}(s)$ determine the fixed transfers for higher pairs.

In equilibrium, higher ability workers acquire lower costs of effort, and higher type occupations invest more to lower exposure to moral hazard. Higher ability physicians work in occupations that have more precise measurement, face stronger incentives, and supply more effort. Productivity increases with type. Because of the complementarity between physicians and occupations, a given increase in physician quality is magnified so that joint surplus increases by a disproportional factor ($\frac{dW^{2,SB}}{d\rho} = (W_c^{2,SB} + W_s^{2,SB}\mu_s^{2,P})c_\rho$). The distribution of earnings across occupations depends on the distribution of ability c, as one would expect, but also on the distribution of measurement risk s. Earning inequalities across occupations increase when there is more heterogeneity in performance risk across occupations.

Efficiency

We now turn to the main implications of the model. Firstly, there is a shortage of effort in all occupations relative to the first best level of effort, because the effective marginal cost of effort, $C_e(1+rs^2C_{ee})$, includes a risk premium due to moral hazard. Most importantly, this shortage of effort increases with occupational risk. As a result, the ratio

between marginal productivity and marginal cost of effort, a measure of occupational efficiency, increases with occupational type.

$$\frac{d}{d\gamma}\frac{\Pi_e(e(\mu^{2,P}(s(\gamma))))}{C_e(e(\mu^{2,P}(s(\gamma)))|\mu^A(s(\gamma)))} = \frac{d}{d\gamma}\frac{1}{b_1(\mu^{2,P}(s(\gamma)),s(\gamma))} > 0 \tag{5}$$

Differences in moral hazard across occupations generates differences in producticvity.

Secondly, some occupations may have to shut down in equilibrium. Participation in the above equilibrium is warranted as long as the surplus of the lowest types is sufficient to cover their reservation utilities

$$W^{2,SB}(c(\rho_0),s(\gamma_0)) - H(c(\rho_0),\rho_0) - K(s(\gamma_0),\gamma_0) \geq -\frac{1}{r}\ln(-U_0) + V_0$$

and this condition holds under the assumption $\ln\left(-U^{0,P}\right) = V^{0,S} = 0$. To further explore the role of outside options, assume that there exists a $\rho^* \in [\rho_0, \rho_1]$ such that

$$W^{2,SB}(c(\rho^*),\mu^{2,S}(c(\rho^*))-H(c(\rho^*),\rho^*)-K(\mu^{2,S}(c(\rho^*),G^{-1}(F(\rho^*))) = -\frac{1}{r}\ln\left(-U^{0,P}\right)+V^{0,S}.$$

All physicians such that $\rho < \rho^*$ and occupations such that $\gamma < G^{-1}(F(\rho^*)) = \gamma^*$ prefer the outside option. Total employment, defined as the mass of employed physicians, $1 - F(\rho^*)$, increases with an improvement in moral hazard technology. Employers with high moral hazard risk have to shut down in equilibrium despite the fact that the marginal productivity of effort is the same in these occupations and in those occupations that do not shut down. The viability of a medical occupation depends on its exposure to moral hazard. The main message of the model is to show that occupations subject to high moral hazard risk will face more difficulties attracting talented physicians.

Thirdly, the only source of inefficiency is due to moral hazard. There do not exist alternative matches or investments that would make any pair better off. In particular, physicians and occupations do not over-invest to improve their match opportunities.

Proposition 2: The equilibrium allocation is constrained Pareto efficient.

The finding that pre-matching investments are constrained efficient extends the analysis of Peters and Siow (2002) to transferable utilities. In our model, the market return functions do not depend only on the matching function, $\mu^{2,\cdot}(.)$, as would be the case under

non-transferable utilities. They also depend on the equilibrium sharing rule. The combination of the equilibrium sharing rules and matching functions give bilaterally efficient investment incentives in period one ($u_c^{2,P}(c) = W_c^2(c, \mu^{2,S}(c))$ and $u_s^{2,S}(s) = W_s^2(\mu^{2,P}(s), s)$). Although the model was presented in the context of a decentralized labor market, all results follow in applications to internal labor markets (e.g. army, professional firms) since Proposition 2 shows that there is no coordination externality.

Forth and finally, the model implies that high talent physicians concentrate in low risk occupations. This does not introduce any distortion relative to the first best allocation, because the sorting of medical occupations and physicians is arbitrary in the absence of moral hazard. But consider an extension of the current model where it is efficient to allocate talent evenly across medical occupations, for example, because of complementarity between different talent levels as in Saint-Paul (2001). The presence of moral hazard would introduce a force that attracts high talent physicians to low risk occupations. This suggests that the equilibrium allocation of talent would be distorted relative to the first best allocation. Although the argument is informal, the model identifies a force that could create a shortage of talent in high risk occupations.

Growth in Specialization

Medical practices with low exposure to moral hazard have an incentive to branch out into independent occupation since by doing so they can attract better physicians and increase joint-surplus. The current model, however, does not deliver this prediction because the set of occupations Γ is fixed by assumption. To make progress, we propose a slightly modified interpretation of the model. Assume that the index γ denotes a medical field instead of a medical occupation. The set of medical fields, Γ, is given and we are interested in how these fields are pooled into medical occupations, where a medical occupation is a set of fields that share the same moral hazard risk. According to that contractual definition of occupation, a measure of specialization is $\Sigma = s(\gamma_0) - s(\gamma_1)$. In the benchmark case where all medical fields face the same exposure to moral hazard ($s(\gamma)$ is constant across γ), this measure says that there is no specialization, $\Sigma = 0$. Specialization increases with Σ.

Under symmetric information, there is no contractual specialization in the sense defined above since medical fields are undifferentiated. Under asymmetric information, however, medical fields get bundled into occupations with identical s. Contractual specialization depends on how exposure to moral hazard varies across fields. To illustrate the possibility of growth in specialization, assume that the moral hazard technology K depends on a shifter denoted by a that captures, for example, developments in scientific technology (e.g. ecography, endoscopy). We have $K(s, \gamma, a)$ where $K_a \leq 0$. Define γ_0 as primary care fields and γ_1 as narrower medical fields and assume furthermore that scientific technology has an impact only in narrow fields, $K_a(s, \gamma_0, a) = 0$. With new developments in scientific technology, $s(\gamma_0)$ remains constant while $s(\gamma_1)$ decreases. As a result, Σ increases and this can be interpreted as an increase in specialization. Growth in specialization is driven by a differential change in the ability to control exposure to moral hazard across fields.

4.3 Example and Discussion

Closed form solutions can be obtained in the case without pre-matching investments. Assortative matching implies $F(\mu^{2,P}(s)) = G(s)$. Contract, effort, and surplus are given by

$$\begin{aligned} b_1(\mu^{2,P}(s)) &= \frac{\pi}{1 + r\mu^{2,P}(s)s^2} \\ e(\mu^{2,P}(s)) &= \frac{\pi}{\mu^{2,P}(s)\left(1 + r\mu^{2,P}(s)s^2\right)} \\ W^{2,SB}(\mu^{2,P}(s), s) &= \frac{\pi^2}{2\mu^{2,P}(s)\left(1 + r\mu^{2,P}(s)s^2\right)} \end{aligned}$$

All three functions decrease with type. Effort and surplus are lower than under the first best allocation.

The main point of the model is to establish the possibility of differential inefficiency across occupations and also of magnification of these inefficiencies through complementarity in matching. To clarify this point, we consider three different scenarios. (1) Assume no moral hazard, and occupations vary in their marginal productivity of effort, $\Pi(e) = \pi e$, where π captures the occupation's type. Then talent varies across occupations, because

of PAM along (c, π), but there is no inefficiency. (2) Introduce moral hazard and assume no heterogeneity across physicians, and occupations again vary in π. Inefficiencies are constant across occupations up to the scale factor π. These first two scenarios show that matching alone and moral hazard alone do not generate differential inefficiencies across occupations. (3) In the case considered in the model, with matching on (c, s) and moral hazard (the same applies to matching over (π, s)), inefficiencies vary across occupations for two reasons. Risk varies across occupations and this is furthermore amplified by the complementarity between talent and risk and the endogenous adjustment of incentives across occupations.

Surplus decreases with type for three reasons: low talent physicians are less productive, work in riskier occupations, and face weaker incentives. Considering the benchmark case with almost no physician heterogeneity makes this point clear. Assume that the support of physician type is $[c_0, c_0 + \varepsilon]$ where ε is a small positive number so that physicians are almost identical. The first best effort level is almost constant, close to $\frac{\pi}{c}$, while the equilibrium effort level decreases as s spans the interval $[s_0, s_1]$. As a result, productivity decreases with type.

Although there exist inefficiencies in all occupations (as long as $s_0 > 0$), the model focuses on relative inefficiencies across occupations. The ratio of the highest to lowest productivity is higher under asymmetric information than under the first best allocation.

$$\frac{W^{2,SB}(c_0, s_0)}{W^{2,SB}(c_1, s_1)} = \frac{c_1}{c_0}\frac{b_1(c_0)}{b_1(c_1)} = \frac{W^{2,FB}(c_0, s_0)}{W^{2,FB}(c_1, s_1)}\frac{1 + rc_1 s_1^2}{1 + rc_0 s_0^2}$$

Distortions relative to the first best allocation are large when s and/or c, and therefore b_1, span a large interval. To further establish that point, we compute the elasticity of surplus to risk.

$$\varepsilon_s^{W^{2,SB}} = \frac{\partial W^{2,SB}}{\partial s}\frac{s}{W^{2,SB}} = -\frac{2}{1-\alpha}\left[1 + (2-\alpha)\varepsilon_s^c\right]$$

where $\varepsilon_s^c = \frac{\partial \mu^{2,P}(s)}{\partial s}\frac{s}{\mu^{2,P}(s)}$ measures the percentage change in risk for a one percent change in physician talent and $\alpha = \frac{b_1}{\pi} \in [0, 1]$ corresponds to the normalized sharing rule. The amplification effect can be large. In fact, $\varepsilon_s^{W^{2,SB}} < -2\left[1 + \varepsilon_s^c\right] < -2$ and a one percent increase in measurement risk implies at least a two percent reduction in surplus. When

the distribution of types are equal up to a constant $F(\rho) = G(\gamma - k)$, we have $\varepsilon_s^c = 1$, and $\varepsilon_s^{W^{2,SB}} < -4$.

A final point on compensation variability is worth mentioning. Paradoxically, compensation risk does not always increase with occupational risk. The variance of compensation is

$$Var\ w = \left(\frac{\pi s}{1 + r\mu^{2,P}(s)s^2} \right)^2$$

A sufficient condition for compensation risk to decrease with occupational risk is $r\mu^{2,P}(s)s^2 < 1$ which is equivalent to $b_1(\mu^{2,P}(s)) < \frac{\pi}{2}$. When the incentive schemes are low powered (the physician gets a share lower than fifty percent), more talented physicians will earn less variable compensation, despite the fact that they face more powerful incentives. This is because they work in less risky occupations. In general, the covariation between occupational risk and pay variability depends on the strength of these countervailing effects.

5 Implications

As mentioned in the introduction, many considerations influence the sorting of physicians and medical occupations. The model, however, focuses exclusively on moral hazard risk. The proper use of the model, therefore, is to consider situations where moral hazard risk varies over time, space, or similar occupations, and to study the impact on sorting, holding other considerations constant. We present three applications along these lines. In addition, the model can be used to formulate normative assessments of policy. To illustrate, we discuss implications of the recent debate on performance measurement.

Generalist versus Specialist Careers

In recent years, medical students tend to favour specialist occupations over being a generalist. Among the factors that influence this decision, DeWitt et al. (1998) report that "subjects cited the ability in specialty practice to have problems 'well-framed,' to 'be the expert,' and to gain mastery over a smaller core of knowledge, as well as the uncertainty inherent in general medicine. Many expressed variations of one physician's opinion that, 'It's easier to be a specialist because there's a smaller area of expertise and one can happily and guiltlessly ignore all other problems'." The model can explain the

recent imbalance of generalists relative to specialists by a differential decrease in risk in the latter occupation, and this interpretation is consistent with the above conclusions. In addition, the finding that more able physicians sort in narrower fields is consistent with the literature. For example, Kiker and Zeh (1998) report that 'It is generally expected that physicians with greater academic ability opt for the more technical specialties over primary care.'

Growth in Specialization

The model also provides an explanation for the recent growth in the number of medical specialties. The recognition of medical specialties started in the late 1920's in an attempt to standardize curriculum, training, and qualification. The number of sub-specialties, measured either by the number of sub-specialties with accredited programs or with certification of individual physicians, has grown from about 30 in the early 1970's to more than 100 in the late 1990's (Donini-Lenhoff, 2000). The appropriate mix of generalist and specialist has been an ongoing topic of debate (Barondess, 2000). Some see specialization as the result of technological and scientific advances and we do not deny that such trends play a role.[16] We argue that in addition to this fragmentation force, the issue of moral hazard may have also played a role in the growth in specialization. The model shows that sub-specialties that cover domains where performance can be assessed more accurately have an incentive to branch out. For example, the sample size effect discussed earlier suggests that moral hazard might be reduced in more specialized fields where physicians repeat the same procedures.

The view presented in this work is consistent with the observation that the growth in specialization is largely decentralized and has been supply-driven. For example, Martini (1993) argues that "the system responds more promptly to professionals' interests and institutionals' service needs" than to "the health need of the population". Some have even argued that the proliferation of specialties diffuses responsibility for clinical care over time and over multiple health disorders which is fully consistent with the view presented in

[16]The growth in sup-specialization in the 90's has occurred in a period where the total number of residents was not increasing (Brotherton et al. 2002), ruling out the hypothesis that scale alone is driving specialization.

our analysis. While generalists are exposed to a common risk associated with unknown ailments, specialists are held responsible only for specific disorders.

A prediction specific to our analysis is that one would expect to observe more sub-specialization when there is more heterogeneity in risk within the set of disorders that belong to a given medical field. In addition, the branching out should be initiated by low risk sub-specialties. This prediction has obvious implications in the context of the malpractice debate. Given that malpractice premia are specialty dependent, those physicians working in low-risk specialties do not want to pool risk with high-risk specialties. We argue that this same force offers a more general explanation for the trend toward specialization.

Malpractice

Kessler et al. (2005) present evidence on the impact of malpractice liability on the supply of physicians. They compared states that adopted legal reforms that limited malpractice liability to those that didn't on trends in physician enrolment, distinguishing specialties with different levels of risk. They considered a wide range of reform to malpractice laws between 1985 and 2001 that affected variables such as the level of damage awards, the possibility for punitive damage, among others. Using a difference in difference approach, they found greater growth in physician supply in states that adopted reforms, and a greater-than-average effect on the supply of physicians in the majority of the high risk specialties.

This evidence is consistent with our model but not definitive. In fact, a decrease in malpractice liability non only decreases risk but also decreases the expected cost of malpractice. Even if physicians were risk-neutral, which would eliminate the force identified in the model, one would expect that malpractice reforms should increase physician supply. Our model predicts that the same response should be observed even after holding the expected cost of malpractice constant. Unfortunately, Kessler et al. did not distinguish the impact of malpractice laws due to changes in expected cost and changes in risk. Using a smaller sample from Arizona, Thornton (2000) holds expected cost constant and finds that malpractice risk has a negative impact on physician supply (a finding consistent with

our analysis) but the effect is not significant.

Even more convincingly, Kessler et al. (2005) showed that the legal reform had a lower impact among physicians practicing in large organizations. Since the expected cost of malpractice should be independent of the size of the organization, the reform should have an impact on the relative supply of physicians only through a change in relative risk. Because it is more difficult to self-insure risk in small organizations, the reform should differentially increase the attractiveness of small organizations, and this prediction is consistent with the evidence presented in Kessler et al.

Policy Implications

Consumer advocacy groups, health insurers, and medical societies share interests in the development of methods that permit to identify and reward better physicians.[17] In a review of physician clinical performance assessment, Landon et al. (2003b) argue that "both patients and health care purchasers desire more effective means of identifying excellent clinicians, and a number of organization have begun discussing and implementing plans of assessing the performance of individual clinicians... Some professional specialty societies have begun encouraging physicians to measure their performance by offering increased recognition to those who participate in voluntary performance assessment." The possibility to reward physician performance is also receiving increasing attention as widespread experimentation is yielding lessons on the impact of pay-for-performance (Armour et al. 2001).[18]

The model suggests that policies geared toward the introduction of performance measurement should take into account the sorting implications of unevenly changing moral hazard risk across medical occupations. An increased emphasis on pay-for-performance, through subsidies to invest in s for example, will change the relative importance of performance incentives across medical occupations, and could increase the importance of PAM, with possible distortions in the allocation of talent. Our model establishes a connection between the debate on performance measurement in medicine and the debate on

[17] See Loeb (2004) for a historical review on the use of performance measurement in health care.

[18] Another approach to increase quality of care is to require medical specialties to administer recertification boards.

the supply physicians across medical occupations, and suggests that in order to internalize externalities across specialties, reforms that influence moral hazard risks should be implemented across-the-board instead of specialty-by-specialty.

6 Conclusions

This paper presents a model of sorting of physicians across medical occupations. Our departing assumption is that it is more difficult to identify physicians' performance in occupations where scientific fact and clinical evidence play a lesser role and where clinical outcomes are uncertain and difficult to compare. The model sheds some light on the possibility of differential inefficiency across medical occupations, the debate on the relative scarcity of talent across occupations, the growth in the number of sub-specialities, and the impact of malpractice risk on career choice. To conclude, we discuss broader applications of the model.

The main message of the model applies to other specialized labor markets. In fact, there are many labor markets where workers have to commit to an occupation and where the ability to measure performance differs across occupations. An occupation where performance cannot be measured precisely could be an occupation where there are no explicit performance measures or more generally an occupation where it takes a long time before individual effort has an impact on organizational performance. This could be because the occupation involves complex tasks, uncertain and changing environments, team work, and other factors that make it difficult to disentangle the role played by different input factors of production and random productivity shocks. As a result, even evaluators who have access to the same objective information (e.g. supervisors, peers, or experts) may disagree about individual performance. Applications include the market for academics (Courty and Marschke, 2008), or a firm's internal labor market with competing career tracks. In these labor markets, our model establishes a relationship across occupations between: (a) exposure to moral hazard risk, (b) the allocation of talent, (c) the use of pay for performance incentives, and (d) productivity differentials.

References

1. Abouleish Amr, Jeffrey Apfelbaum, Donald Prough, John Williams, Jay Roskoph, William Johnston, and Charles Whitten. (2005). "The Prevalence and Characteristics of Incentive Plans for Clinical Productivity Among Academic Anesthesiology Programs." Anesthesia Analgesia. 100:493-501

2. Ackerberg, D. and M. Botticini (2002), "Endogenous Matching and the Empirical Determinants of Contract Form," Journal of Political Economy, 110:564-591.

3. Alchian, Armen, and Harold Demsetz. 1972. "Production, Information costs and Economic Organization." American Economic Review. 62, 777-795.

4. Angell, M. and P. Kassirer. (1996). "Quality and the Medical Marketplace — Following Elephants." New England Journal of Medicine. 335:883-885.

5. Armour, Brian, Melinda Pitts, Ross Maclean, Charles Cangialose, Mark Kishel, Hirohisa Imai, Jeff Etchason. (2001) "The Effect of Explicit Financial Incentives on Physician Behavior," Archives of Internal Medicine. 161:1261-1266

6. Arrow, Kenneth. (1963). "Uncertainty and the Welfare Economics of Medical Care." American Economic Review. 53:941-973.

7. Barondess, Jeremiah. (2000). "Specialization and the Physician Workforce: Drivers and Determinants." Journal of the American Medical Association. 284: 1299-1301.

8. Besley T., Ghatak M. (2005). "Competition and Incentives with Motivated Agents." The American Economic Review. 95, 616-636

9. Bland, Kirby I., George Isaacs. "Contemporary Trends in Student Selection of Medical Specialties: The Potential Impact on General Surgery." (2002) Archives of Surgery. 137:259-267.

10. Brotherton, Sarah, Paul H. Rockey, Sylvia I. Etzel. (2005) "US Graduate Medical Education, 2004-2005 Trends in Primary Care Specialties." Journal of the American Medical Association. 294:1075-1082.

11. Chiappori, P.A. and B. Salanié (2003), "Testing Contract Theory: A Survey of Some Recent Work", in M. Dewatripont, L. Hansen and S. Turnovsky, eds., Advances in Economics and Econometrics, Cambridge University Press, Cambridge.

12. Committee on Quality Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington D.C.: Institute of Medicine, National Academy Press; 2001.

13. Courty, Pascal and Gerald Marschke. (2008) "Incentives in Academica", Mimeo, EUI.

14. DeWitt, Dawn, Randall Curtis, and Wylie Burke. "What Influences Career Choices Among Graduates of a Primary Care Training Program?". Journal of General Internal Medecine. 1998 April; 13(4): 257–261.

15. Donini-Lenhoff, Fred, Hannah Hedrick. (2000) "Growth of Specialization in Graduate Medical Education." Journal of the American Medical Association. 284:1284-1289.

16. Dorsey, Ray, David Jarjoura, Gregory Rutecki. (2003) "Influence of Controllable Lifestyle on Recent Trends in Specialty Choice by US Medical Students." Journal of American Medical Association. 290, 1173-1178.

17. Dranove, David, Daniel Kessler, Mark McClellan and Mark Satterthwaite. (2003). "Is more Information Better? The Effects of Report Cards on Cardiovascular Providers and Consumers." Journal of Political Economy. 11: 555-588.

18. Ferris et al. (2007) "Physician Specialty Societies And The Development Of Physician Performance Measures." Health Affairs. 26: 1712-1719.

19. Gaynor, Martin, and Paul Gertler. (1995). "Moral Hazard and Risk Spreading in Partnerships." The RAND Journal of Economics. 4:591-613.

20. Gold, Marsha. (1999). "Financial incentives. Current Realities and Challenges for Physicians." Journal of General Internal Medicine. 14: S6-S12.

21. Hojat, Mohammadreza, Joseph Gonnella, Thomas Nasca, Salvatore Mangione, Michael Vergare, and Michael Magee. (2002) "Physician Empathy: Definition, Components, Measurement, and Relationship to Gender and Specialty." American Journal of Psychiatry. 159, 1563-1569.

22. Holmstrom, Bengt & Milgrom, Paul. (1994) "The Firm as an Incentive System." American Economic Review. 84, 972-91.

23. Holmstrom, B., and Milgrom, P. (1991) "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." The Journal of Law, Economics, and Organization. 7, 24-52.

24. Kessler, Daniel, and William Sage, David Becker. (2005) "Impact of Malpractice Reforms on the Supply of Physician Service." Journal of the American Medical Association. 293, 2618-2625.

25. Kiker, B.F., and Michael Zeh. (1998). "Relative Income Expectation, Expected Malpractice Premium Costs, and Other Determinants of Physician Specialty Choice." Journal of Health and Social Behavior. 39. 152-67.

26. Landon, Bruce, James Reschovsky, David Blumenthal. (2003a) "Changes in Career Satisfaction Among Primary Care and Specialist Physicians, 1997-2001." Journal of the American Medical Association . 289, 442-449.

27. Landon, Bruce, Sharon-Lise Normand, David Blumenthal, and Jennifer Daley. (2003b) "Physician Clinical Performance Assessment: Prospects and Barriers." Journal of American Medical Association. 290, 1183-1189.

28. Leffler, Keith. (1978). "Physician Licensure: Competition and Monopoly in American Medicine." Journal of Law and Economics. 21:165—186.

29. Loeb, Jerod M. (2004). "The Current State of Performance Measurement in Health Care." International Journal for Quality in Health Care. 16:i5-i9.

30. Martini, Carlos. (1992). "Graduate Medical Education in the Changing Environment of Medecine." The Journal of the American Medical Association. 268, 1097-1105.

31. Nicholson, Sean. (2002) "Physician Specialty Choice Under Uncertainty," Journal of Labor Economics. 20, 816-47.

32. Peters, Michael and Aloysius Siow. (2002) "Competing Pre-marital Investments." Journal of. Political Economy, 110, 592–608.

33. Prendergast, Canice. (1999) "The Provision of Incentives in Firms." Journal of Economic Literature. 37, 7-63.

34. Prendergast, Canice. (2002) "The Tenuous Trade-Off between Risk and Incentives." Journal of Political Economy. 110. 1071-102.

35. Roth, Alvin. (2008). "What Have we Learned from Market Design?" Hahn Lecture. Economic Journal. 118. 285-310..

36. Roth, Alvin and Marilda Sotomayor. 1990. Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis, Cambridge, Cambridge University Press.

37. Saint-Paul, Gilles. (2001) "On the Distribution of Income and Worker Assignment under Intra-Firm Spillovers, with an Application to Ideas and Networks." Journal of Political Economy. 109, 1-37.

38. Serfes, Konstantinos. (2005). "Risk Sharing vs. Incentives: Contract design under Two-Sided Heterogeneity." Economics Letters, 88, 343-349.

39. Serfes, Konstantinos. (forthcoming). "Endogenous Matching in a Market with Heterogeneous Principals and Agents." International Journal of Game Theory.

40. Thornton, James and Fred Esposto. (2002) "How Important Are Economic Factors in Choice of Medical Specialty?" Health Economics. 12, 67 - 73.

41. Thornton, James. (2000). "Physician Choice of Medical Specialty: Do Economic Incentives Matter." Applied Economics. 32. 1419-28.

42. Weeks, W. and A. Wallace (2002) "Long-Term Financial Implications of Specialty Training for Physicians." The American Journal of Medicine. 113, 393-399.

Appendix: Proofs

For the sake of completeness, we present all the steps to derive the equilibrium.

Equilibrium Definition

We define the continuation payoffs (in certainty equivalent units) in period two, conditional on matching and investments as $U^{2,P}(b,c,s)$ and $U^{2,S}(b,c,s)$. Using the effort rule $e(b_1,c)$ from equation (1), simple computations give $U^{2,P}(b,c,s) = b_0 + b_1 e(b_1,c) - C(e(b_1,c)|c) - \frac{r}{2}(sC_e(e(b_1,c)|c))^2$ for $s \neq \emptyset$ and $U^{2,S}(b,c,s) = \Pi(e(b_1,c)) - b_0 - b_1 e(b_1,c)$ for $c \neq \emptyset$.[19] The equilibrium conditions can be formally stated as:

(1) The investment rules are defined as $c(\rho) = ArgMax_c \left(u^{2,P}(c) - H(c|\rho)\right)$ and $s(\gamma) = ArgMax_s \left(u^{2,S}(s) - K(s|\gamma)\right)$.

(2) Stability is satisfied if physician c such that $\mu^{2,S}(c) \neq \emptyset$ does not want to deviate from $B(c)$ and $\mu^{2,S}(c)$ in stage two

$$B(c), \mu^{2,S}(c) \in ArgMax_{b,s\neq\emptyset}\ U^{2,P}(b,c,s) \quad s.t.\ U^{2,S}(b,c,s) \geq u^{2,S}(s)$$

and specialty s such that $\mu^{2,P}(s) \neq \emptyset$ does not want to deviate from $B(\mu^{2,P}(s))$and $\mu^{2,P}(s)$

$$B(\mu^{2,P}(s)), \mu^{2,P}(s) \in ArgMax_{b,c\neq\emptyset}\ U^{2,S}(b,c,s) \quad s.t.\ U^{2,P}(b,c,s) \geq u^{2,P}(c)\ .$$

(3) In addition, worker ρ is willing to participate in period one, $-exp[-r(u^{2,P}(c(\rho)) - H(c(\rho)|\rho))] \geq U^0$ and the same holds for specialty s, $u^{2,S}(s(\gamma)) - K(s(\gamma)|\gamma)) \geq U^{0,S}$.

(4) Rational expectations hold if $u^{2,P}(c) = U^{2,P}(B(c), c, \mu^{2,S}(c))$ and $u^{2,S}(s) = U^{2,S}(B(\mu^{2,P}(s)), \mu^{2,S}(s)$

Since physicians and occupations can select the outside option in period one, we do not have to reconsider this option in period two.

Proof of Lemma 1

The proof goes by contradiction. Assume (c,s) are matched and work under contract $b = (b_0, b_1)$ such that $b_1 \neq b_1(c)$. Stability implies that

$$U^{2,S}(b,c,s) \geq Max_{b'}\ U^{2,S}(b',c,s) \quad s.t.\ U^{2,P}(b',c,s) \geq u^{2,P}(c)$$

The maximum computed under the restriction that the constraint binds is weakly dominated by the maximum without this restriction.

$$U^{2,S}(b,c,s) \geq Max_{b'}\ U^{2,S}(b',c,s) \quad s.t.\ U^{2,P}(b',c,s) = u^{2,P}(c)$$

The restriction that the constraint binds can be expressed as

$$b'_0 + b'_1 e(b'_1,c) - C(e(b'_1,c)|c) - \frac{r}{2}(sb'_1)^2 = b_0 + b_1 e(b_1,c) - C(e(b_1,c)|c) - \frac{r}{2}(sb_1)^2$$

[19] Holmstrom and Milgrom (1991), p. 179.

Plugging the above equality in the objective function and cancelling terms gives

$$\Pi(e(b_1,c)) - C(e(b_1,c)|c) - \frac{r}{2}(sC_e(e(b_1,c)|c))^2 \geq$$
$$Max_{b'}\left(\Pi(e(b',c)) - C(e(b',c)|c) - \frac{r}{2}(sC_e(e(b',c)|c))^2\right)$$

The maximization problem on the right hand side has a unique optimum as long as $C_{ee}^2 + C_e C_{eee} > 0$ which holds under A1a. The optimum is achieved at $b_1(c)$. The above inequality contradicts the assumption that $b_1 \neq b_1(c)$. QED

Restatement of the Period Two Matching Problem
Denote by $W^2(c,s)$ the period two certainty equivalent of pair (c,s).

$$W^{2,SB}(c,s) = Max_b\ \left\{U^{2,P}(b,c,s) + U^{2,S}(b,c,s)\right\} = Max_e\ \left\{\Pi(e) - C(e|c) - \frac{r}{2}(sC_e(e|c))^2\right\}$$

Assumption A1a is sufficient to guarantee that there is a unique maximum. We rewrite the stability conditions in period two for any pair (c,s) as

$$u^{2,P}(c) + u^{2,S}(\mu^{2,S}(c)) = W^{2,SB}(c,\mu^{2,S}(c)) \text{ for any } c \text{ such that } \mu^{2,S}(c) \neq \emptyset$$
$$u^{2,P}(c) \geq W^{2,SB}(c,s) - u^{2,S}(s) \text{ for any } c,s$$

The first conditions say that any matched pair splits their joint surplus. The second condition corresponds to the stability conditions that no physician or specialty would be better off in a different match.

Proof of Lemma 2
The proof proceeds in three steps.

Claim 1: $W_{cs}^{2,SB}(c,s) > 0$.

The cross derivative can be written as

$$W_{cs}^{2,SB} = -\frac{2rsc_e\left(\Pi_{ee}c_{ec} + rs^2c_e(c_{ece} - c_{eee})\right)}{\Pi_{ee} - c_{ee} - rs^2(c_e c_{eee} + c_{ee}^2)}$$

which is positive under A1.

Claim 2: In any equilibrium, there is PAM in (c,s) *in period two.*

The proof follows by contradiction. Assume $c_1 > c_0$ and $s_1 > s_0$ and pairs (c_1,s_0) and (c_0,s_1) are matched. Since s_0 does not deviate to c_0 and s_1 does not deviate to c_1, we have

$$\begin{aligned} W^{2,SB}(c_1,s_0) - u^{2,P}(c_1) &\geq W^{2,SB}(c_0,s_0) - u^{2,P}(c_0) \\ W^{2,SB}(c_0,s_1) - u^{2,P}(c_0) &\geq W^{2,SB}(c_1,s_1) - u^{2,P}(c_1) \end{aligned}$$

Summing up these two inequalities give

$$W^{2,SB}(c_0,s_1) + W^{2,SB}(c_1,s_0) \geq W^{2,SB}(c_1,s_1) + W^{2,SB}(c_0,s_0)$$

which contradicts claim 1 stating that c and s are complement in $W^{2,SB}$.

Claim 3: In any equilibrium, there is PAM in $(c, -\gamma)$.

The proof again follows by contradiction. Assume $c_1 > c_0$ and $\gamma_1 > \gamma_0$ and pairs (c_0, γ_0) and (c_1, γ_1) are matched. Two cases can be distinguished. The case $s(\gamma_1) < s(\gamma_0)$ leads to a contradiction with claim 2 stating that there is PAM in (c, s). Consider next the possibility that $s(\gamma_1) > s(\gamma_0)$. In period one, γ_0 does not want to mimic γ_1 and γ_1 does not want to mimic γ_0. This implies

$$\begin{aligned} u^{2,S}(s(\gamma_0)) &\geq u^{2,S}(s(\gamma_1)) + (K(s(\gamma_1)|\gamma_1) - K(s(\gamma_1)|\gamma_0)) \\ u^{2,S}(s(\gamma_1)) &\geq u^{2,S}(s(\gamma_0)) + (K(s(\gamma_0)|\gamma_0) - K(s(\gamma_0)|\gamma_1)) \end{aligned}$$

Summing up these two inequalities gives

$$K(s(\gamma_1)|\gamma_1) - K(s(\gamma_0)|\gamma_1) \leq K(s(\gamma_1)|\gamma_0) - K(s(\gamma_0)|\gamma_0)$$

which contradicts the fact that γ and s are complement in K.

To conclude, note that a similar proof as the one presented in claim 3 shows that there is also PAM in $(-\rho, s)$. Lemma 2 then follows by putting together PAM in $(c, -\gamma)$, (c, s), and $(-\rho, s)$. QED

Proof of Proposition 1

Lemma 2 says that in there is PAM in (c, s) in any equilibrium. We first compute the market return functions under period two equilibrium matching, then the equilibrium investments in period one, and finally show that the equilibrium investments in period one are consistent with the market return functions.

Claim 1: Assume there is a continuum of types (c, s) *in period two. There exists a unique equilibrium (up to constants* $u^{2,P}(c_1)$ *and* $u^{2,S}(s_1)$*) and it satisfies PAM. The market return functions* $u^{2,P}(.)$ *and* $u^{2,S}(.)$ *are given by*

$$\begin{aligned} u^{2,S}(s) &= u^{2,S}(s_1) - \int_s^{s_1} W_s^{2,SB}(\mu^{2,P}(s'), s')ds' \\ u^{2,P}(c) &= u^{2,P}(c_1) - \int_c^{c_1} W_c^{2,SB}(c', \mu^{2,S}(c'))dc' \end{aligned}$$

and $u^{2,P}(c_1) + u^{2,S}(s_1) = W^{2,SB}(c_1, s_1)$.

Lemma 2 shows that PAM is the only candidate equilibrium. To show existence, we first show that the above payoff functions satisfy stability. Consider the possibility that type c's deviates and match with s. The maximum increase in utility c can get is

$$\begin{gathered} W^{2,SB}(c, s) - u^{2,S}(s) - u^{2,P}(c) = \\ W^{2,SB}(c, s) - W^{2,SB}(c, \mu^{2,S}(c)) + u^{2,S}(\mu^{2,S}(c)) - u^{2,S}(s) = \\ \int_{\mu^{2,S}(c)}^{s} \left(W_s^{2,SB}(c, s') - W_s^{2,SB}(\mu^{2,P}(s'), s') \right) ds' \leq 0 \end{gathered}$$

Therefore, physician c does not deviate. The same argument applies to specialty s.

To show uniqueness, note that in period two, physician c has to prefer $\mu^{2,P}(c)$ over any other specialty, implying $u_c^{2,P}(c) = W_c^{2,SB}(c, \mu^{2,S}(c))$. Similarly, we have $u_s^{2,S}(s) = W_s^{2,SB}(\mu^{2,P}(s), s)$. These two differential equations determine the functions $u^{2,P}$ and $u^{2,S}$ up to integration constants $(u_1^{2,P}, u_1^{2,S})$ which have to satisfy $u_1^{2,P} + u_1^{2,S} = W^{2,SB}(c_1, s_1)$.

Claim 2: Equilibrium investments satisfy (4).

In period one, physician ρ maximizes $u^{2,P}(c) - H(c|\rho)$. The first order condition to the investment problem gives

$$u_c^{2,P}(c) - H_c(c|\rho) = 0.$$

In equilibrium, $u_c^{2,P}(c) = W_c^{2,SB}(c, \mu^{2,S}(c))$ and after replacement, we obtain the first equation in (4). The second equation can be similarly obtained by solving the specialty's investment problem. The second order condition to investment problems are satisfied if $(H_{cc} - W_{cc}^2)(K_{ss} - W_{ss}^{2,SB}) > W_{sc}^{2,SB}$ which holds under A2b.

Under A2, there exists a unique solution (c, s) to the system

$$\begin{cases} H_c(c|\rho) = W_c^{2,SB}(c, s) \\ K_s(s|\gamma) = W_s^{2,SB}(c, s) \end{cases}$$

Therefore, the functions $(c(\rho), s(\gamma))$ are uniquely determined by (4).

Claim 3: Participation holds as long as $-exp[-r(u^{2,P}(c_1) - H(c(\rho_1)|\rho_1))] \geq U^{0,P}$ and $u^{2,S}(s_1) - K(s(\gamma_1)|\gamma_1)) \geq U^{0,S}$.

Type ρ_0 participates in period one if $u^{2,P}(c_1)$ is such that $-exp[-r(u^{2,P}(c_1) - H(c(\rho_1)|\rho_1))] \geq U^{0,P}$. Types $\rho > \rho_0$ also participates because expected period one utility is increasing in ρ. Similarly, γ_1 participates in period if $u^{2,S}(s_1)$ is such that $u^{2,S}(s_1) - K(s(\gamma_1)|\gamma_1)) \geq U^{0,S}$ and higher types also participates because period one utility is increasing in γ.

Claim 4: Monotonicity of investment rules.

The final step is to check that period two matching defined by (3) is consistent with the investment rules defined by (4). This will be the case if $c(\rho)$ and $s(\gamma)$, are monotonously decreasing in type. To show that this is the case, rewrite (4) as a function of γ. PAM in period one implies that γ is matched with a physician denoted $\mu^{1,P}(\gamma)$ such that

$$F(\mu^{1,P}(\gamma)) = G(\gamma)$$

In addition, we have

$$\mu^{2,P}(s(\gamma)) = c(\mu^{1,P}(\gamma))$$

Replacing these expressions in (4) gives

$$\begin{cases} H_c(c(\mu^{1,P}(\gamma)), \mu^{1,S}(\gamma)) = W_c^{2,SB}(c(\mu^{1,P}(\gamma)), s(\gamma)) \\ K_s(s(\gamma), \gamma) = W_s^{2,SB}(c(\mu^{1,P}(\gamma)), s(\gamma)) \end{cases}.$$

Taking full derivative in the above system, we have

$$\begin{bmatrix} (W_{cc}^{2,SB} - H_{cc})\mu_{\gamma}^{1,P} & W_{sc}^{2,SB} \\ W_{sc}^{2,SB}\mu_{\gamma}^{1,P} & (W_{cc}^{2,SB} - H_{cc}) \end{bmatrix} \begin{pmatrix} c_{\rho} \\ s_{\gamma} \end{pmatrix} = \begin{pmatrix} H_{c\rho}\mu_{\gamma}^{1,P} \\ K_{s\gamma} \end{pmatrix}.$$

Since $\mu^{1,P} > 0$, A2 is a sufficient condition for monotonicity, $c_{\rho} < 0$ and $s_{\gamma} < 0$. To conclude, note that monotonicity of the investment rules implies that there is a continuum of types (c, s) in period two and matching takes place according to (3). QED

Proof of Proposition 2

Given that pair (ρ, γ) is matched together, the social planner sets the investments to maximize the period one joint surplus $W^{2,SB}(c, s) - H(c|\rho) - K(s|\gamma)$. The information constrained period one surplus of pair (ρ, γ) measured in certainty equivalent units is

$$W^1(\rho, \gamma) = Max_{c,s}\{W^{2,SB}(c, s) - H(c|\rho) - K(s|\gamma)\}.$$

The social planner selects a matching rule in period one that maximizes the joint surplus $W^1(\rho, \gamma)$. Since ρ and γ are complement in the function $W^1(\rho, \gamma)$, PAM in (ρ, γ) is efficient. The investment rules that maximize $W^{2,SB}(c, s) - H(c|\rho) - K(s|\gamma)$ under PAM are monotonic in (ρ, γ) and correspond to the equilibrium investment rules. The constrained Pareto efficient allocation is identical to the equilibrium allocation. QED