On the Sorting of Physicians across Medical Specialties: Understanding Shortages and Growth in Specialization

Pascal Courty and Gerald Marschke [1,2]

January 15, 2008

ABSTRACT: We model the sorting of medical students across specialties and identify a mechanism that explains the possibility of differential shortage across specialties. The model combines moral hazard and matching of physicians and specialties with pre-matching investments. In equilibrium assortative matching takes place; more able physicians join specialties where performance is measured more precisely, face more powerful performance incentives, and are more productive. Under-consumption of health services relative to the first best allocation increases with specialty risk. Specialties with risk above a given threshold have to shut down. The model offers an explanation for recent trends, such as the differential shortage of physicians across specialties, the differential impact of malpractice risk across specialties, and the growth in the number of sub-specialties.

KEYWORDS: Performance risk, incentives, matching, pre-matching investment, shortage, specialization, medical specialty.

J.E.L: D82, I10, J31, J33, L23.

[1]Pascal Courty: European University Institute, Via della Piazzuola 43, 50133, Firenze, Italy. Gerald Marschke: Department of Economics and Department of Public Administration and Policy, BA-110, University at Albany, State University of New York, Albany, NY 12222.

[2]Comments welcome at pcourty@eui.eu. We would like to thank David Perez Castrillo, Javier Rivas and participants at EUI and ASSET. All opinions and any errors are ours.

1 Introduction

The distribution of medical students' career choice among specialties is the object of intense debates. It is generally acknowledged, for example, that there is a chronic deficit of physicians in general medicine and in some specialties such as psychiatry (Brotherton et al. 2005). More recently, the decrease in the proportion of medical students who choose a career in general surgery has raised concerns (Barshes et al. 2004).[1] Governments and medical organizations sometimes intervene with a variety of incentives and regulations (Thornton and Esposto, 2002). Why is there a shortage of physicians for some specialties? If wages can freely adjust, as they often do in many countries, how can chronic deficits persist?

This paper develops a theoretical framework to capture how medical students sort across specialties and identifies a mechanism that explains the possibility of differential shortage across specialties. The model combines moral hazard (Holmstrom and Milgrom, 1987) and matching of physicians and specialties with pre-matching investments (Peters and Siow, 2002). A key feature of the model is that when medical students select a specialty, they take into account how their performance will be evaluated in their future career. We present much evidence in the next section that the ability to measure performance varies greatly across specialties. It is harder to identify and reward excellence in specialties where decision-making is less grounded in scientific fact and clinical evidence and where clinical outcomes are uncertain and difficult to compare. As a result, the cost of using performance incentives is higher in specialties with greater performance risk.

We show that in equilibrium assortative matching takes place. More able physicians join specialties where performance is measured more precisely, face more powerful performance incentives, and are more productive. Even when all specialties are identical ex-ante in terms of marginal productivity of effort, productivity is higher in less risky specialties. Two forces drive this result. First, physicians with lower cost of effort end up in occupations with more precise performance measurement. Second, the incentive

[1] Approximately 7.8% of graduating US medical students chose general surgery in 1987 compared with only 5.8% in 2002 (Barshes et al., 2004).

scheme is endogenous in our setup and this feature 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 specialties where risk is too high have to shut down in equilibrium. They would be able to attract only low productivity physicians, and would have to set low-powered incentives, so that the overall surplus would not cover outside options. A broader interpretation of this result is that high risk specialties face greater difficulties attracting physicians. Secondly, those physicians who accept a position in a high-risk specialty face less powerful incentives and supply less effort. These two channels imply that inefficiencies increase with specialty risk. In addition, this inefficiency differential across specialties increases as the distributions of physician talent and specialty 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 doctor. The distribution of consumption across specialties, however, is distorted relative to the first best one. There is under-consumption of high risk services. The differential between marginal productivity and marginal cost of effort increases with occupational risk, and we interpret this outcome as a shortage of service in high risk specialties.

We recognize that many factors influence the sorting of physicians across specialties. For example, lifestyle and work schedule have been shown to influence career choice (Landon et al. 2003a). Some of these factors also explain why the relative demand for some specialties 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 but this cannot explain the excess residency positions in some specialties. The main contribution of this work is to identify a mechanism that explains distortions in the distribution of health care consumption across specialties, relative to the first best allocation.

After presenting the main results, we discuss several implications. The analysis sug-

gests an explanation for the shortage of generalists relative to specialists. Assume that performance risk has decreased in specialty careers relative to generalist ones, which is consistent with evidence presented in DeWitt et al (1998). The implication is that specialty careers would become more attractive in relative terms.

In addition, we argue that the model provides an explanation for the growth specialization over the past decades. The number of sub-specialties has grown from about 30 in the early 70's to more than 100 in the late 90's (Donini-Lenhoff, 2000). 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, that differentially affects specialties, will have implications for the relative shortage of talent across specialties.

We extend the model to pre-matching investments. Physicians and specialties can invest resources to respectively lower cost of effort and measurement risk. We show that the equilibrium is constrained Pareto efficient in the sense that there do not exist alternative matches or investments that would make any subset of pairs better off; the only source of inefficiency is due to moral hazard. The finding that pre-matching investments are constrained efficient complements the analysis of Peters and Siow (2002), who assume, in the context of an application to a marriage markets, that utilities are non-transferable. In contrast, we assume transferable utility. Physicians and specialties have rational expectations about the return from pre-matching investments and the market return functions provide investment incentives that are bilaterally efficient.

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 measurement risk (Prendergast, 2002). Little attention has been dedicated to study broader implications of performance measurement at a more macro level, across occupations or within a special-

ized labor market, as we do in this paper. In particular, there is to our knowledge no work investigating the possibility that the absence of good performance measures (objective or subjective) may lead to a failure to organize an economic activity.[2] Our central assumption that performance measurement heterogeneity across occupation may influence matching, plays an important role in the recent empirical literature studying the relation between incentive and 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 focus on differentiated occupations and consider 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, the 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 (1987) 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.

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.

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

2 Medical Specialties, Career Choice, and Performance Risk

About one-third of physicians are generalists or "primary care" doctors. There are specialties within primary care such as internal medicine and pediatrics. When patients' specific health needs require further treatment, generalist physicians send them to see a specialist physician. Specialist physicians differ from generalists in that they focus on treating a particular system or part of the body, such as neurologists who study the brain, or cardiologists who study the heart. 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 agency). Recently, medical societies have also started to help develop performance measures and pay for performance schemes (Ferris et al. 2007).

Career Choice and Shortage

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 generalist to specialist physicians 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, a balanced specialty mix of medical care in the long-term. For example, in 1993 and 1994, the Physician Payment Review Commission recommended Congress to implement a system of quantitative restrictions on positions.

There is a large literature studying the choice of medical specialty both in medical sciences (e.g. Weeks and Wallace (2002) and economics (Nicholson, 2002). A large body of research has shown that demographic characteristics influence career choice, suggesting that there exist preferences over specialties. In addition, there is also much evidence that economic incentives matter. Among other considerations, perceived future earn-

ings, 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 why shortages occur. To single out the driving force in our mechanism for shortages, the model will assume that all specialties are identical in all respects except in the level of measurement risk, and that physicians select a medical specialty only on the basis of expected future utility.

Performance Risk

Another departing point of the model is that performance risk varies across specialties. More specifically, the model follows Holmstron and Milgrom (1987) and assumes that the risk imposed on physicians increases with the level of incentive. In addition, the model also assumes that this incentive-induced risk varies across specialties. To fix ideas, we discuss different sources of risk that are consistent with this view.

The ability to measure performance varies greatly across medical specialties. There are many reasons for this. The information available on outcomes of care and clinical processes depends on the specialty. 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." 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. Landon et al. (2003b) conclude that "few medical specialties have an evidence base that is robust and comprehensive enough to support physician clinical performance assessment." 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."[3][4]

The model also assumes that performance risk can affect physicians' career outcomes. This could happen through several channels. Traditionally, reputation and word-of-mouth, as well as specialty disciplining boards, have provided feedback loops between physician performance and physician reward. More recently, a number of private firms and public organizations have started to compile information on individual physicians' performance and are making it available over the Internet. The National Committee for Quality Assurance, a widely recognized non-profit organization dedicated to improving health care quality, helps patients to identify high performing physicians in their state. Similarly, HealthGrades is a health care rating organization that covers hospitals, nursing homes and physicians. One would expect ratings to influence decisions by patients and managed care organizations, and therefore physician demand. Most importantly for our study, rankings should be less reliable, and possibly less widespread as well, in specialties where it is more difficult to measure performance.

In addition some risks, such as malpractice risk for example, depend on the specialty. Anesthesiology, radiology, and surgery, for example, are more exposed to malpractice risk and this is reflected in higher insurance ratings (Kessler et al. 2005). Malpractice risk also depends on the legal environment which varies across states. The framework adopted in this paper applies to the extent that malpractice is subject to moral hazard.[5]

As mentioned above, medical societies try to influence the amount of risk associated with a given specialty. Landon et al (2003b) report that "Some professional specialty

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

[4]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). Empathy is difficult to measure and this may explain why the ambulatory care performance measure recommendations from the AQA Performance Measurement Workgroup, which forms the basis of pay-for-performance by third party payers such as Medicare, contain no measure of empathetic attitude.

[5]Physicians in high malpractice specialties may be more likely to work in larger practices where risk can be shared, but also where one would expect weaker performance incentives, which is consistent with our argument.

societies have begun encouraging physicians to measure their performance by offering increased recognition to those who participate in voluntary performance assessment." To reduce the number of actors in the analysis, the model assumes that physicians are hired by specialties. By specialty, we will mean both the unit where the physician is employed and also the medical society that controls the specialty. This abstraction is meant to capture the point that specialties are competing for talent although strictly speaking, physicians do not work for specialties.[6]

3 Model

The objective of the model is to identify a mechanism that can generate differential shortage of physicians across specialties and also to reveal the factors that generate this differential. For this reason, we selectively include in the model the features that can generate such differential or magnify it. Because these features are not necessarily present simultaneously, the shortages 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). The other components of the model rest on general functional forms. Following the assortative matching literature, we model matching using unidimensional preference ordering. This narrows attention on the main force that can generate the effect we are interested in. A more realistic model would acknowledge the fact that matching takes place along other dimensions but this would not change the nature of the results.

There are three periods. In period one, physicians and specialties invest in human capital and monitoring respectively. At the end period one, the distribution of human capital and monitoring investments are observed. In the second period, physicians and specialties match and agree on a contract. In the third period, physicians exert effort, nature draws performance, and contracts are executed.

[6] An essential assumption of the model is that physicians share performance risk. This is realisitic for physicians that work in health maintenance organizations, in hospitals, or in group practices.

There is a unit continuum of physicians indexed by $\rho \in R = [\rho_0, \rho_1]$. Physician type ρ is distributed with density $f > 0$ and distribution F, where F is continuous, $F(\rho_0) = 0$ and $F(\rho_1) = 1$. 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 $H(c|\rho) > 0$ where $H_c < 0$, $H_{cc} > 0$, $H_\rho < 0$, and $H_{\rho c} > 0$. The utility of a physician of type ρ who selects cost of effort c, exerts efforts e, and is paid wage w is

$$U(e, c, w|\rho) = -exp[-r(w - C(e|c) - H(c|\rho))].$$

There is a unit continuum of medical specialties, indexed by γ, which are taken as given. γ is distributed according to density $g > 0$ and distribution G, which is continuous over $\Gamma = [\gamma_0, \gamma_1]$, and such that $G(\gamma_0) = 0$ and $G(\gamma_1) = 1$. Work effort is subject to moral hazard. Each specialty can only observes an imperfect measure of effort. A specialty, however, can invest to increase the precision of this measure. Specialty γ can achieve monitoring risk $s \geq 0$ at cost $K(s|\gamma) > 0$ where $K_s < 0$, $K_{ss} > 0$, $K_\rho < 0$, and $K_{s\rho} > 0$. Investments in monitoring should be interpreted broadly. It could capture the investment made by the unit or hospital hiring the physician. Alternatively, medical specialty societies also invest in monitoring quality through re-licensure, disciplining, and development of performance measurement. Performance measurement in specialty with risk s is

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

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 specialties.

In period two, physicians and specialties 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 specialty rewards the physician according to the agreed rule $w(m)$. We first assume that all specialties equally value $\Pi(e)$ effort level e such that $\Pi' > 0$ and $\Pi'' < 0$. We later discuss the case of heterogeneous productivity across specialties. Specialties maximize profits, $\Pi(e) - Ew(m) - K(s|\gamma)$, or

$$\Pi(e) - b_1 e - b_0 - K(s|\gamma).$$

We focus on stable matching (Roth and Sotomayor, 1990). We denote $\mu^A(s)$ (resp. $\mu^P(c)$) the physician (resp. specialty) matched with specialty s (resp. physician c) such that $\mu^A(\mu^P(c)) = c$ if specialty s (resp. physician c) is matched and $\mu^A(s) = \emptyset$ (resp. $\mu^P(c) = \emptyset$) otherwise. A contract function associates a contract $B(c) = (b_0(c), b_1(c))$ to each matched pair. The outside option of specialties and physicians who have not matched are U^0 and V^0 respectively. In stage two, we denote $v(s)$ the expected payoff of specialty s and $u(c)$ the certainty equivalent continuation payoff of physician c. (c is indifferent between receiving $u(c)$ for sure and matching with $\mu^P(c)$ under contract $B(c)$. As will become clear soon, CARA utility implies that the certainty equivalent does not depend on physician type ρ.) Following Peter and Siow (200), we define a rational expectation equilibrium as:

(1) A set of investment rules $c(\rho)$ and $s(\gamma)$ for physicians and specialties that maximize their payoffs conditional on expectations about $u()$ and $v()$.

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

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

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

(4) Physicians and specialties have rational expectations: the functions $u()$ and $v()$ are consistent with $\mu^P(c)$, $B(c)$, and $e(c)$.

The functions $u(c)$ and $v(s)$ correspond to the market return of investments. As in

Peters and Siow (2002), physicians and specialties choose optimal investments given their expectations about the market returns. The main difference is that utility is transferable in our model so the functions $u(c)$ and $v(s)$ 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 $(c(\rho), s(\gamma), \mu^P(c), e(c), B(c))$. The main innovation of the model is to capture the fact that incentive risk varies across specialties and to allow for matching. For the sake of generality, we have introduced the possibility of pre-matching investments in monitoring, and this addresses the concern that the quality of monitoring can be 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 inessential.

In addition to boundary conditions, $C_e(0|c) = 0$, $C_e(\infty|c) = \infty$, and similarly for H_c and K_s, two technical conditions are sufficient to demonstrate equilibrium uniqueness.

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 $W(c,s) = Max_e \left\{\Pi(e) - C(e|c) - \frac{r}{2}(sC_e(e|c))^2\right\}$ the period two information constrained joint surplus function in certainty equivalent units. The meaning of this expression will become clear after Lemma 1. A1b is sufficient to show that effort and monitoring are complement in the joint surplus function.

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

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 $\Pi(e) = \pi e$. 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). Then specialties do not invest in monitoring, sorting is arbitrary, and a 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 specialty where she is employed. Since specialties are identical, they receive a constant payoff, which is determined such that both sides of the market are willing to participate. Physicians receive the residual surplus.

In the application with no pre-matching investments and quadratic cost of effort, the effort supplied by physician c is

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

Let $W^1(c, s)$ represent the period one surplus produced by pair (c, s) measured in monetary terms,[7]

$$W^{1,FB}(c, s) = \frac{\pi^2}{2c}.$$

Surplus increases in talent and is independent of specialty risk.

4.2 Asymmetric Information

A formal derivation of the equilibrium is presented in the appendix. We analyze the problem backward. Consider a physician of type c who has matched in period two with specialty s and agreed to contract (b_0, b_1). In period 3, the physician sets e to maximize $b_1 e - C(e|c)$. The period 3 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.

[7]This is the sum of the specialty profit and physician certainty equivalent and the later is equal to the physician monetary payoff in the abscence of uncertainty.

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 shows 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-temporal separability meaning that we can solve for matching and contracting in stage two independently of the stage one pre-matching investment choices.

The definition of $W(c, s)$ becomes clear. In period two, physician c and occupation s agree on contract $b_1(c, s)$ and the certainty equivalent continuation payoff is $W(c, s)$. This corresponds to the maximum payoff (in certainty equivalent units) that the pair can achieve under incentive compatibility. We have $W(c, s) = u(c) + v(s)$.

We now turn to the matching problem. To start, we assume that all physicians and specialties match. A sufficient condition for this to hold is $U_0 = V_0 = H(0|\rho_0) = K(0|\gamma_0) = 0$. We can now state our main result that matching is positive assortative (PAM).

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 specialty risk are complement in the joint surplus function $W(c, s)$. This alone implies PAM in (c, s) in period two. Second, investments in lower cost of effort and 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) which is the source of efficiency differential across specialties, as we will see soon. 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. The main point is that the analysis is robust to pre-matching investments and the results are magnified under complementarity.

Cost of effort and specialty risk are complement in $W(c, s)$ when the sharing rule $b_1(c, s)$ is endogenously determined. When $b_1(c, s)$ is exogenously given, complementarity disappears ($W_{cs} = 0$) and assortative matching does not follow. It would still be the case that specialties with high risk would be less attractive, but the magnifying effects due to matching and endogenous incentive would disappear. Therefore, the model applies primarily to those physicians working in health maintenance organizations, hospitals, and group practices, and secondarily to all physicians to the extent that malpractice risk sharing varies across specialties.

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. More generally, 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 firms differ in riskiness s as we do in this paper) and shows that it is possible to characterize the equilibrium only in specific cases.[8] In contrast, we consider matching between worker ability and occupational risk. Since c and s are complement in W only PAM can occur.

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 worker

[8]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 and PAM (Negative AM) holds. He cannot characterize the equilibrium for intermediate ranges of rs^2.

would have identical cost function but would differ in term of marginal productivity (worker 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 $b_0(c(\rho_1))$. Period two matching is defined by

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

contracting is defined by (2) and investments by

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

Equations (3) and (4) define the matching function and pre-matching investment. The sharing rule is defined by (2). The stability conditions define the period two continuation payoffs (up to a constant) according to $u_c(c) = W_c(c, \mu^P(c))$ and $v_s(s) = W_s(\mu^P(s), s)$ and the constant is determined by the allocation of surplus between the lowest pair $b_0(c(\rho_1))$ which can take any value such that the participation constraints of the lowest types are satisfied. The resulting $u(c)$ and $v(s)$ determine the fixed transfer for higher pairs.

In equilibrium, higher ability workers acquire lower costs of effort, and higher type specialties acquire more precise monitoring technologies. Higher ability physicians work in specialties that have more precise measurement, face stronger incentives, and supply more effort. Productivity increases with type. Because of the complementarity between physicians and specialties, a given increase in the quality of physician (or specialty monitoring technology) is magnified so that the surplus generated by physician-specialty pair increases by a disproportional factor ($\frac{dW}{d\rho} = (W_c + W_s\mu_s^P)c_\rho$).

Participation in the above equilibrium is warranted as long as surplus of the lowest types is sufficient to cover their reservation utilities

$$W(c(\rho_1), s(\gamma_1)) - H(c(\rho_1), \rho_1) - K(s(\gamma_1), \gamma_1) \geq \ln U_0 + V_0$$

and this condition holds under the assumption $U_0 = V_0 = H(0|\rho_0) = K(0|\gamma_0) = 0$. To further explore the role of outside options, assume that there exists a $\rho^* < \rho_1$ such that

$$W(c(\rho^*), \mu^p(c(\rho^*)) - H(c(\rho^*), \rho^*) - K(\mu^p(c(\rho^*), G^{-1}(F(\rho^*))) = \ln U_0 + V_0.$$

All physicians such that $\rho < \rho^*$ and specialties such that $\gamma < G^{-1}(F(\rho^*)) = \gamma^*$ prefer the outside option. Specialties with poor measurement technologies are shut down in equilibrium despite the fact that the marginal productivity of effort is the same in these specialties and in those specialties that are not shut down. The viability of a specialty depends on the existence of reliable performance measures. Total employment, defined as the mass of employed physicians, $F(\rho^*)$, increases with an improvement in monitoring technology for those specialties below the marginal one $G^{-1}(F(\rho^*))$.

Shortage may be given two interpretations in the context of the model. First, shortage may happen in an extreme sense. Strictly speaking, high risk specialties are shut down but the main message of the model is to show that specialties that have poor measurement technologies will face more difficulties attracting physicians. Second, the extent of inefficiency varies across specialties. The differential between the marginal productivity and marginal cost of effort increases with specialty risk. In this sense, there is a shortage of effort in high risk specialties.

In equilibrium, high talent physicians work in low risk specialties. This does not introduce any distortion relative to the first best allocation, because the sorting of specialties 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 specialties, 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 in low risk specialties. This suggests that the equilibrium allocation would distort the allocation of talent 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 specialties relative to the first best allocation.

Proposition 2: The equilibrium allocation is constrained Pareto efficient.

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. A matched pair of physician and specialty bilaterally internalizes the gains from investments. In addition, physicians and specialties do not over invest to improve their match opportunities. The finding that pre-matching investments are constrained efficient extends the analysis of Peters and Siow (2002) to transferable utilities. Specialties and physicians select the constrained efficient level of effort cost c and risk s because they fully internalize the benefit of marginal investment under the equilibrium market return functions, $u_c(c) = W_c(c, \mu^P(c))$ and $v_s(s) = W_s(\mu^A(s), s)$. In our model, the market return functions do not depend only on the matching function, $\mu(.)$, 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 efficient investment incentives in period one.

4.3 Example and Discussion

In the case without pre-matching investments, assortative matching implies $F(\mu^A(s)) = G(s)$. Contract, effort, and surplus are given by

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

Incentives, effort, and surplus 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 specialties and also of magnification of these inefficiencies through complementarity in matching and pre-matching investments. To clarify this point, we consider three different scenarios. (1) Assume no moral hazard, and specialties vary in their marginal productivity of effort, $\Pi(e) = \pi e$, where π captures the specialty type. Then talent varies across specialties, because of PAM along (c, π), but there is no inefficiency. (2) Introduce

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

Surplus decreases with type for three reasons: low talent physicians are less productive, work in riskier specialties, 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 $[c_0, c_0 + \zeta]$ where ζ 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 specialties (as long as $s_0 > 0$), the model focuses on the relative inefficiency across specialties. The ratio of the highest to lowest productivity is higher under asymmetric information than under the first best allocation.

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

Distortions relative to the first best allocation are large when s and/or c, and therefore b_1, span a large interval. While variations in s capture the differential role of risk across specialties, variations in c amplify the role of risk through assortative matching.

To further establish that point, we compute the elasticity of surplus to risk.

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

where $\varepsilon_s^c = \frac{d\mu^A(s)}{ds}\frac{s}{\mu^A(s)}$ 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 percentage sharing rule. The amplification effect can be large. In fact, $\varepsilon_s^{W^1} < -2\left[1 + \varepsilon_s^c\right] < -2$ and a one percent increase in measurement noise implies at least a two percent decrease 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 $\varepsilon_s^{W^1} < -4$. The multiplicative factor emerges both because of assortative matching and because of the endogeneity of the incentive scheme. Better workers match with better measurement technologies and face more powerful incentives. The former effect takes place in markets with matching and complementarity (e.g. Rosen's superstar model (1981)) while the later effect is specific to this model.

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

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

A sufficient condition for compensation risk to decrease with specialty risk is $r\mu^A(s)s^2 < 1$ which is equivalent to $b_1 < \frac{\pi}{2}$. 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 specialties. In general, the covariation between specialty risk and pay variability depends on the strength of these countervailing effects. This suggests that one has to be careful measuring risk empirically. Pay variability cannot be used as a proxy for specialty risk.

5 Implications

As mentioned in the introduction, many considerations influence the sorting of physicians across specialties. The model, however, focuses exclusively on performance risk. The proper use of the model, therefore, is to consider situations where performance risk varies over time, space, or similar occupations, and to study the impact on sorting, holding other considerations constant. We present two applications along these lines. In addition, the model can be used to make normative assessment of policy. For example, we discuss implications of the recent emphasis on performance measurement.

Generalist Shortage

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 shortage of generalists relative to specialists by a differential decrease in risk in the latter occupation, and this interpretation is consistent with the above conclusions.

Growth in Specialization

The model also provides an explanation for the growth in the number of medical specialties. The recognition of medical specialties started in the late 1920's in an attempt to standardize curriculum format, 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.[9] We argue that in addition to this fragmentation force, the issue of performance risk 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.

This interpretation 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 our analysis. While generalists are exposed to a common risk associated with unknown

[9]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 is driving sub-specialization.

ailments, specialists are held responsible only for specific disorders.

A prediction specific to our analysis is that one would expect to observe more specialization in domains where there is more heterogeneity in the risks associated with different disorders. 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 used variations in malpractice premia across specialties to identify five high-risk specialties. 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 three of the five 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.

They showed, however, that the reform had a greater impact among physicians practicing in non-group settings (excluding physicians working in health maintenance organizations, hospital, or in public sector practices). Since physicians working in non-group

settings face the same expected cost of malpractice as those working in group settings, the reform should have an impact on the relative supply of physicians only through a change in relative risk. The reform should reduce risk more in non-group settings because these physicians cannot pool risk as well as those physicians working in group settings. Our analysis says that such a decrease in relative risk should differentially increase the attractiveness of non-group settings 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.[10] In a review of physician clinical performance assessment, Landon et al. (2003) 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." 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). Another approach to increase quality of care is to require medical specialties to administer re-certification boards.

The model suggests that policies geared toward the introduction of performance measurement should take into account the sorting implications of unevenly changing performance risk across specialties. An increased role of performance risk tends to increase the importance of PAM (with distortions in the allocation of talent) and to disproportionately reduce the role of incentives in high risk specialties. As a result, an increased emphasis on pay-for-performance will have implications for the supply of physicians across specialties.

6 Conclusions

This paper presents a model of sorting of physicians across medical specialties. Our departing assumption is that it is more difficult to identify better physicians in specialties

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

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 debate about the relative scarcity of talent across specialties, 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 both within the context of our medical labor market application and also to other specialized labor markets.

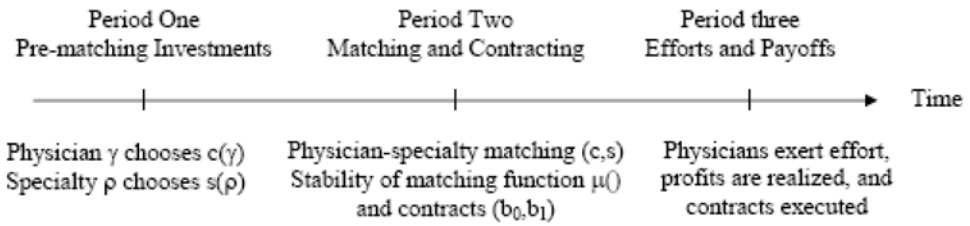
The main message of the model applies to other specialized labor markets. In fact, there are many labor markets where the ability to measure performance differs across occupations. An occupation where performance cannot be measured precisely could be an occupation where there are no explicit performance 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 who have to select a field, or a firm's internal labor market with competing career tracks (Courty and Marschke, 2008). In these labor markets, our model establishes a relationship across occupations between: (a) the availability of reliable measures of output, (b) the type of workers employed, and (c) the use of pay for performance incentives. More able workers are more likely to work in occupations where performance can be measured more precisely, face more powerful incentives, exert more effort, and are more productive.

The model could be applied to other fields of medicine. For example, our analysis could be applied to the debate between preventive versus curative medicine. In the US, expenditures on prevention included in the national health account represented only 3 percent of total health expenditure in 1988.[11] In 2004, the number of residents in the US in training for preventive medicine specialties represented only 0.4 percent of all residents.

[11]This share across 22 OECD countries varies from 0.6 percent (Italy) to 8 percent (Canada) (Bekker-Grob et al. 2007).

Our model suggests that preventive medicine may not be able to attract talent, and to grow relative to curative medicine, because it is difficult to measure the effectiveness of preventive treatments.

Figure 1: Timeline



References

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

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

3. 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

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

5. Barshes, N., A. Vavra, A. Miller, F. Brunicardi, J. Goss, J. Sweeney. (2004) "General surgery as a career: A contemporary review of factors central to medical student specialty choice." Journal of the American College of Surgeons. 199, 792-799.

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

7. 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.

8. 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.

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

10. 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.

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

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

13. 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.

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

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. Loeb, Jerod M. (2004). "The current state of performance measurement in health care." International Journal for Quality in Health Care. 16:i5-i9.

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

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

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

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

25. Rosen, Sherwin. 1981. "The Economics of Superstars." American Economic Review. 71, 845-58.

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

27. 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.

28. Serfes, Konstantinos. (2005). "Risk sharing vs. incentives: Contract design under two-sided heterogeneity." Economics Letters, 88, 343-349.

29. Serfes, Konstantinos. (forthcoming). "Endogenous matching in a market with heterogeneous principals and agents." International Journal of Game Theory.

30. Thornton, James and Fred Esposto. (2002) "How important are economic factors in choice of medical specialty?" Health Economics. 12, 67 - 73.

31. 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, keeping in mind that parts of the argument are standard.

Definition of the Equilibrium

We compute the continuation payoffs in period two (in certainty equivalent units), using the effort rule $e(b_1, c)$ from equation (1) in period three, as $U^2(b, c, s) = b_0 + b_1 e(b_1, c) - C(e(b_1, c)|c) - \frac{r}{2}(sC_e(e(b_1, c)|c))^2$ (Holmstrom and Milgrom (1991), p. 179) for $s \neq \emptyset$ and $V^2(b, c, s) = \Pi(e(b_1, c)) - b_0 - b_1 e(b_1, c)$ for $c \neq \emptyset$. The equilibrium conditions can be formally stated as:

(1) The investment rules are defined as $c(\rho) = ArgMax_c\,(u(c) - H(c|\rho))$ and $s(\gamma) = ArgMax_s\,(v(s) - K(s|\gamma))$.

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

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

and specialty s such that $\mu^A(s) \neq \emptyset$ does not want to deviate from $B(\mu^A(s))$and $\mu^A(s)$

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

.

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

(4) Rational expectations hold if $u(c) = U^2(B(c), c, \mu^P(c))$ and $v(s) = V^2(B(\mu^A(s)), \mu^A(s), s)$.

Since physicians and specialties 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

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

The maximum computed under the restriction that the constraint binds has to be weakly dominated by the maximum without this restriction.

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

The restriction that the constraint binds can be expressed as

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

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 $b_1(c,s)$. The above inequality contradicts the assumption that $b_1 \neq b_1(c)$. QED

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

$$W(c,s) = Max_e\ \left\{\Pi(e) - C(e|c) - \frac{r}{2}(sC_e(e|c))^2\right\}$$

We rewrite the stability conditions in period two for any pair (c,s) as

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

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: $W_{cs}(c,s) > 0$.

The cross derivative can be written as

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

which is positive under A1.

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

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

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

Summing up these two inequalities give

$$W(c_0,s_1) + W(c_1,s_0) \geq W(c_1,s_1) + W(c_0,s_0)$$

which contradicts claim 1 stating that c and s are complement in W.

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

$$\begin{aligned} v(s(\gamma_0)) &\geq v(s(\gamma_1)) + (K(s(\gamma_1)|\gamma_1) - K(s(\gamma_1)|\gamma_0)) \\ v(s(\gamma_1)) &\geq v(s(\gamma_0)) + (K(s(\gamma_0)|\gamma_0) - K(s(\gamma_0)|\gamma_1)) \end{aligned}$$

Summing up these two inequalities gives

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

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 functionx 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 2. There exists a unique equilibrium and it satisfies PAM. The market return functions* $u(.)$ *and* $v(.)$ *are given by*

$$\begin{aligned} v(s) &= v_1 - \int_s^{s_1} W_s(\mu^A(s'), s')ds' \\ u(c) &= u_1 - \int_c^{c_1} W_c(c', \mu^P(c'))dc' \end{aligned}$$

and $u_1 + v_1 = W(c_1, s_1)$ *where the level of* u_1 *and* v_1 *can be arbitrarily set to meet the participation constraints of the highest types in period one.*

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

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

Therefore, physician c does not deviate. The same argument applies to specialty s. Next, note that all participation constraints are satisfied because utility is decreasing in type and the highest type is willing to participate by assumption.

To show uniqueness, note that in period two, physician c has to prefer $\mu^P(c)$ over any other specialty, implying $u_c(c) = W_c(c, \mu^P(c))$. Similarly, we have $v_s(s) = W_s(\mu^A(s), s)$. These two differential equations determine the functions u and v up to integration constants (u_1, v_1) which have to satisfy $u_1 + v_1 = W(c_1, s_1)$.

Claim 2: Equilibrium investments satisfy (4).

In period 1, physician ρ maximizes $u(c) - H(c|\rho)$. The first order condition to the investment problem gives

$$u_c(c) - H_c(c|\rho) = 0.$$

In equilibrium, $u_c(c) = W_c(c, \mu^P(c))$ 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 $(H_{cc} - W_{cc})(K_{ss} - W_{ss}) > W_{sc}^2$ which holds under A2b.

Since A2 is sufficient to guarantee that there exists a unique solution (c, s) to the system

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

the set of equations in (4) have a unique solution $(c(\rho), s(\gamma))$.

Participation in period one holds if $-exp[-r(u_1 - H(c(\rho_1)|\rho_1))] \geq U^0$ and $v_1 - K(s(\gamma_1)|\gamma_1)) \geq V^0$.

Claim 3: Monotonicity of investment rules.

The final step is to check that period two matching is defined by (3) and the investment rules are 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 $\mu^{1A}(\gamma)$ such that

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

In addition, we have

$$\mu^A(s(\gamma)) = c(\mu^{1A}(\gamma))$$

Replacing these expressions in (4) gives

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

Taking full derivative in the above system gives

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

Since $\mu_{\gamma}^{1A} > 0$, 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 $\mu^{A}()$. 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 $W(c, s) - H(c|\rho) - K(s|\gamma)$. The information constrained surplus of pair (ρ, γ) measured in certainty equivalent units is

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

The social planner selects a matching rule in period one that maximinzes the joint surplus $W^{1}(\rho, \gamma)$. Since the function $W^{1}(\rho, \gamma)$ is supermodular, PAM in (ρ, γ) is efficient. The investment rules that maximize $W(c, s) - H(c|\rho) - K(s|\gamma)$ 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