Can We Measure Hospital Quality from Physicians' Choices?

Matilde P. Machado, Ricardo Mora, Antonio Romero-Medina
Universidad Carlos III de Madrid[1,2]

May 5, 2008

[1]Keywords: Hospital Quality, Ranking, Nested Logit, Revealed Preference, Physicians labor market.
[2]We acknowledge the support of the Ramón Areces Foundation (III Concurso Nacional). We thank Pedro Albarrán, James Banks, Lola Collado, Michael Creel, Zvi Eckstein, Rachel Griffith, Maia Güell, and Marcos Vera for valuable comments. We also thank Elvira González from the Ministerio de Sanidad y Consumo for providing the main data, Adelheid Holl for providing us with the travel time between cities data, Josep Vaqué and the Sociedad Española de Medicina Preventiva, Salud Pública e Higiene as well as all hospitals that gave us consent to use their confidential data on nosocomial infections. We are also thankful to Mercedes Cabañas, Oscar Alberto Véliz and Laura Muñóz who helped us with the data collection. This article was supported by an unrestricted educational grant awarded jointly to the Universities Carlos III de Madrid and Pompeu Fabra de Barcelona by The Merck Foundation, the philanthropic arm of Merck Co. Inc., White House Station, New Jersey, USA.

Abstract

In this paper, we propose an alternative methodology to rank hospitals based on the choices of Medical Schools graduates over training vacancies.

We exploit the physicians' hospital choices to infer quality differentials among hospitals. Our methodology for measuring relative hospital quality has the following desirable properties: a) robust to manipulation from the hospital's administration; b) conditional on having enough observations, it can be extended to allow for differences in quality across specialties within a hospital; c) inexpensive in terms of data requirements, d) not subject to selection bias from patients nor hospital screening of patients; and e) unlike other rankings based on experts' evaluations, it does not require physicians to provide a complete ranking of all hospitals. We apply our methodology to the Spanish case and find, among other results, the following. First, the probability of choosing the best hospital relative to the worst hospital is statistically significantly different from zero. Second, physicians value proximity and nearby hospitals are seen as more substitutable. Third, observable time-invariant city characteristics are unrelated to results. Finally, our estimates for physicians' hospital valuations are significantly correlated to more traditional hospital quality measures.

1 Introduction

Assessing the quality of health care providers is a priority of many governments, state agencies, insurance companies and often also patients. Accurate measures of provider's quality can be used to establish incentive mechanisms and to identify which providers need quality improvements. In addition, when made public, these measures can increase market transparency and competition. Public report cards in certain US states and the NHS ranking system in the UK are two attempts at constructing quality rankings of health care providers. Although the need for such rankings is widely recognized, the number of criticisms at these attempts reveal the difficulties involved in this task (see for example the review by Shahian et al., 2001 and the references therein). Most criticisms alert to the inadequate risk-adjustment and the potential for strategic behavior among providers, such as patient selection by hospitals (Dranove et al., 2003, Moscucci et al., 2005), as well as hospital selection by patients based on poor information[1]. Nonetheless, the available studies show that there is a wide variation in the performance of health care providers (e.g. Shahian et al., 2001 and Burgess et al., 2003) and that, in some cases, differences in providers have widened in the past decades (e.g. McLellan and Staiger, 1999a).

Hospital quality assessment is particularly complex due to several reasons: First, hospitals produce a wide range of heterogeneous services, which makes it impossible to define "hospital production" in a simple way (see, for instance, McClellan and Staiger, 1999b, Maxwell, 1984). Second, the randomness of hospital output, e.g. because of small number of patients or uncommon conditions, and the existence of confounding factors, such as location-specific patient health characteristics, may lead to noisy measures of hospital quality. Third, patient selection and other non-random sources of patients assignment (see, for instance, Gowrisankaran and Town, 1999 and Geweke, Gowrisankaran and Town, 2003) will bias estimates of hospital quality. Fourth, due to data collection costs, there is a lack of follow-up measures of treatment results. Finally, in many countries, such as Spain, there is neither public hospital rankings nor public data on hospital output measures such as mortality rates. In this paper, we provide a methodology to rank hospitals which is not affected by these problems because it is not based on hospital output measures but is based on publicly available data on physicians' labour market. More specifically, we use data on the choices over hospital vacancies made by young physicians, hereafter referred to as physicians, at the beginning of their specialized training. Our methodology is based on a revealed preference argument similar to the one used in Avery et al. (2005) for US colleges.

The validity of our methodology to rank hospitals lies in three assumptions. First, physicians should aim to get a vacancy in those hospitals that offer the best training, possibly after controlling for personal factors; Second, physicians are well informed decision makers who are qualified to assess hospital training quality; Third, higher quality hospitals provide on average higher quality training. Given these three assumptions, we argue that hospital quality differencials can be inferred from physicians choices of training vacancies.

The first assumption is reasonable in this very competitive and tight labor market where the place of

[1] An interesting summary of report cards databases can be found in http://www.uihealthcare.com/infofor/corporaterelations/rep

training provides a signal to the post-training market.[2] The second assumption relies on the importance that this decision entails for the physicians' career prospects. On the one hand, training usually lasts between 3 to 5 years and is an essential component of the physicians' formation, affecting decisively their future income. On the other hand, the relative costs of gathering information on hospital training quality seem to be relatively low within the physicians' network.[3] The third assumption can be relaxed, in which case our ranking would simply reflect hospital training quality. Hospital quality stems from the quality of hospital staff, and equipment. Hence, our third assumption implies that staff and/or equipment that contributes to higher hospital quality also provide better training on average. If our three assumptions hold, then we would expect a high correlation between hospital quality and our ranking based on the physician's choices. We confirm this prediction by computing this correlation in a small subsample of hospitals for which more traditional quality measures (i.e., aggregate mortality rates and prevalence of nosocomial infections) exist.

We adapt our methodology to the Spanish physicians labor market, which has interesting properties. In Spain, after graduation from medical school, physicians willing to start a career as specialists must pass a national exam. Conditional on passing the exam, they choose hospital training vacancies sequentially depending on their average grade.[4] The sequential structure of the problem implies dependence across observations. As physicians gradually fill in the vacancies, they decrease the choice sets of remaining physicians. We propose a model similar in spirit to the sequence of "exploded" conditional multinomial logit proposed by Bradlow and Fader (2001). Given our assumptions, the physicians who choose first are likely to grab the best vacancies while the ones who choose last are stuck with the worst vacancies available. Naturally, physicians' choices also depend both on personal considerations. For example, when choosing a hospital, physicians are simultaneously choosing place of residence for as long as the training period. To account for these effects, we complement our data with location-specific characteristics. Our model is then extended, following McFadden's nested logit, to account for different patterns of substitutability across hospitals in different locations.

In short, we exploit the physicians' sequence of hospital choices to infer quality differentials among hospitals. Our methodology for measuring relative hospital quality has the following desirable properties: a) robust to manipulation from the hospital's administration; b) conditional on having enough observations, it can be extended to allow for differences in quality across specialties within a hospital; c) inexpensive in terms of data requirements, d) not subject to selection bias from patients (Gowrisankaran and Town, 1999) nor hospital screening of patients; and e) unlike other rankings based on experts' evaluations, it does not require physicians to provide a complete ranking of all hospitals.

We expect property a) to hold because it should not be possible for hospitals to influence physicians' choices but through their performance. Besides hospital quality, one could also claim that physicians

[2] Specialist certificates, obtained after completing the training period, specify the hospital where the training took place. The name of the training hospital is, therefore, used in the labor market as a signal of quality.

[3] For example, private agencies publish guides with information to help physicians with their vacancy choices.

[4] More precisely, the average grade is obtained as a weighted average of the exam grade (75%) and the physician's GPA (25%).

worry about tenure promotion. Physicians might tradeoff quality for less future job uncertainty. Hospitals administrations could give informal assurances of future promotions to attract good candidates.[5] Since these promises are unobservable to the econometrician, a way to control for this effect would be to proxy for future openings with a variable such as the number of tenured physicians per number of beds. Unfortunately, this data are not available. Instead, in our regressions, we control for number of hospital beds, and the number of physicians per 1000 inhabitants at the local level. Property b) follows because hospitals open vacancies in several specialties. Property c) holds because physicians' choices are routinely collected. With respect to property d), our relative quality measure is not subject to selection bias because it is neither based on patients' nor on hospitals' decisions. As argued by Dranove et al. (2003), report cards based on perfectly adjusted mortality rates would still cause hospital selection of patients leading to biased inferences of hospital quality. Finally, physicians only need local information, i.e. physicians gather costly information on the hospitals which are more likely to be among the best in their expected choice set given their rank. This local information that our methodology exploits is likely more precise than, for example, if they were asked to provide a complete ranking of all hospitals.

We find that the probability of choosing the best hospital relative to the worst hospital is statistically significantly different from zero. In our different model specifications, this odds ratio ranges from 55.7 to 166.8. In all specifications, we also control for interactions between gender and specialty dummies. We find significant differences in the way different specialties are valued by male and female physicians. Nevertheless, our results suggest that the sequential algorithm by which vacancies are chosen precludes neither male nor female physicians from choosing according to their preferences. Specifications differ in two dimensions: the inclusion of hospital characteristics and the treatment of location preferences. Most hospital characteristics are not significant, and their inclusion has little effect on the rankings of hospitals. In contrast, the estimates for the location parameters are very precisely estimated and adding complexity to the modelling of location affects the ranking by more than the inclusion of hospital controls. We find that physicians value proximity and that nearby hospitals are seen as more substitutable. Overall, our rankings are very correlated across different model specifications.

To evaluate our claim that physicians' choices are related to hospital quality, we perform two types of checks. First, it could be argued that in our framework, time invariant city characteristics cannot be separately identified from hospital quality since they contribute to the attractiveness of hospitals in the city. We show that observable time-invariant city characteristics are unrelated to results. Moreover, to control for unobservable time invariant city characteristics that contribute to the attractiveness of hospitals, we also construct within-city rankings. Our results show that city-specific unobservable effects are unimportant. Second, we test whether physicians' choices are related to other traditional measures of hospital quality. Specifically, we compare our rankings with two alternative indicators of hospital quality that we were able to gather for a sub-sample of hospitals: the risk-adjusted prevalence of nosocomial infections and the mortality rates. The results of these tests corroborate that our estimates for physicians' hospital valuations are significantly correlated to more traditional hospital quality

[5] Here we should add that in order to get a tenure position at a hospital, physicians have to go through a national contest – Antonio is working on the details.

measures, which supports the validity of our methodology to construct a ranking of hospital quality.

The remainder of this paper proceeds as follows. Section 2 describes the literature on hospital evaluation. Section 3 sketches the institutional background in Spain. Section 4 describes the data set, the definition of variables. Section 5 presents the econometric framework. Section 6 presents our estimation results. Section 7 discusses and compares our ranking results with rankings based on alternative measures of hospital quality. Section 8 provides some concluding comments. Finally, the Appendix contains figures and tables.

2 Literature Review

Early evaluations of hospital production were based on accounting exercises of the inputs used during an average stay. As better data were made available, other hospital indicators, such as the length of the average stay or the average turnover, were gradually incorporated in the studies. Yet, it soon became clear that the use of such indicators could be easily influenced by hospital administrators so that improvements in the indicators were often associated with situations for which hospital quality had not actually improved. This was considered a serious drawback for the practical usefulness of these indicators; Ideally, quality measures should not be easily manipulated by the hospital administrators (see for instance, Lu, 1999, for a description of this problem in the context of substance abuse treatment).

More recently, some studies have used treatment results, such as hospitals' disease-specific mortality rates and nosocomial infection rates (Gaynes, 1997[6]), to assess hospital quality after controlling for observable characteristics. For instance, Normand *et al.* (1997) compare mortality rates for acute myocardial infarction after using hospital and patient characteristics as risk-adjusters.

When using treatment results, two difficulties may still arise. First, since the patient's true health status is not reflected in just one binary indicator, such as a imminent survival, the results are necessarily a partial analysis of hospital quality. Recent studies have tried to address this limitation by incorporating additional health measures over time (see, for instance, McClellan and Staiger, 1999a, 1999b, and Ackerberg, Machado and Riordan, 2006). Nonetheless, the potential for biased estimates of hospital quality due to insufficient gathering of relevant health information remains.[7] The obvious solution to this problem, i.e. the exhaustive gathering of information regarding treatment results, does not seem practical as data collection costs could increase substantially.

The second difficulty with simple risk-adjustment is that it fails to adjust for non-observable factors. For example, Gowrisankaran and Town (1999) as well as McLellan and Staiger (1999a) argue that

[6] Nosocomial infections are defined as infections which are neither present nor in incubation at the moment of admission into a hospital. We will describe the literature that uses nosocomial infection rates in detail in the Section 7.

[7] McLellan and Staiger (1999a) find that hospital characteristics such as teaching institution status, for-profit organization status, number of beds, and volume cannot explain all the differences in standardized mortality rates. This finding suggests that unobservables may explain them.

the endogeneity in the process of patients' assignment to hospitals is a source of bias when estimating hospital quality because the best hospitals will usually be assigned the most severe cases. Gowrisankaran and Town propose a simple linear instrumental variables approach to the mortality rate model using distances to non-selected hospitals as instruments of hospital choice to deal with the endogeneity but, they caution, their method is not appropriate to construct a ranking of hospitals due to the high standard deviations of the estimated hospital-dummy coefficients . McLellan and Staiger, on the other hand, decide to restrict their analyses to heart attack patients who have less of a possibility to choose hospitals due to the urgency of treatment. Geweke, Gowrisankaran and Town (2003) go one step further and jointly model the hospital choice and the mortality outcome for Medicare patients suffering from pneumonia. They compute the probability of mortality under the hypothetical experiment of random admission to hospitals using bayesian Markov Chain Monte Carlo simulation techniques. In a somewhat different vein, Ackerberg et. al. (2006) address the problem of endogeneity in their study on alcoholism treatment by distinguishing between the intrinsic quality of the health provider and the effect of the patient's unobservable characteristics in the assignment to health providers.

In contrast to the previous studies where patient level data are used, we use public data on physicians' choices over training vacancies at the beginning of their careers. Thus, our proposal relies on a revealed preference argument in the labor market for training vacancies. Avery et al. (2005) use a similar argument to rank colleges in the United States. Our methodologies and applications, however, differ in several ways. First, Avery et al. (2005) aknowledge a potential self-selection problem that would bias their ranking results. In their application, students' choice sets are not exogenous, instead they are a subset of colleges pre-selected by the students. In our application, choice sets are given for each physician. Second, in their data, students' residence is exogenous and precisely recorded. As we argue in Section 5.2, information on physicians' residence in our dataset is measured with error and this error is potentially correlated with physician's ability. We control for location by using college location, which is precisely measured but may be endogenous. We address the endogeneity issue by estimating a nested logit model in constrast to the multinomial logit estimated by Avery et al. (2005).

3 The labor market for specialists

In Spain, it is mandatory for medical school graduates to complete training programs in hospitals during a number of years before they can practice as specialist physicians. Not all hospitals open training vacancies. Hospitals must fullfil certain requirements before they can open training vacancies. These requirements are established by the Ministry of Education and the Ministry of Health. The number of vacancies in each hospital is determined by the government after consultation with hospital administrators and depends on the hospital training capacity and the needs of the population in the surrounding area. In conversations with government officials involved in the process we were told hospitals seek to obtain as many vacancies as they can manage as an unexpensive way of increasing staff. The data suggest that given hospital size, the number of vacancies is unrelated to hospital quality.[8]

Overall, according to data from the National Catalogue of Hospitals, only around 22% of health care providers offer training programs for specialists. These providers share a number of features. All of them are hospitals and, on average, larger than those institutions which do not train specialists. Training hospitals have, on average, 540 beds as opposed to an average of 117 beds for those hospitals without training vacancies. In addition, training hospitals have special health care equipment, such as extracorporeal shock wave lithotripsy, cobalt treatment equipment, and particle accelerators. Nearly 75% of training hospitals report at least one type of special equipment while the figure for non-training hospitals is only 36%. The percentage of private and public hospitals that open training vacancies are 4.65% and 43%, respectively. These statistics suggest that our sample includes a large proportion of the most important hospitals in the Spanish health care system.

Hospitals open vacancies in different specialties. Across all hospitals, the total number of vacancies per specialty-year is very unequal.

The match between physicians and vacancies follows a serial dictator procedure. First, physicians are ranked according to an average score, which is obtained as a weighted average between a standardized national exam score[9] —with a weight of 75%— and the medical school grade-point-average — with a weight of 25%. Every year during our sample period, the number of physicians who pass the exam is higher than the total number of vacancies available.[10] Physicians may choose a vacancy following the sequence established by the rank. Some physicians, however, decide to drop out from the process. Most of these have a low rank, which indicates that they prefer an outside option to the vacancies still available to them (e.g. Gonzalez, 2004). Conditional on choosing a vacancy, physicians' dominant strategy is to choose their best option from those still available.[11]

Wages for training positions are the same across all hospitals. Hence, differences in nominal salary should not influence physicians' choices. In order to control for real wage differences we include housing prices in some of our regression specifications.

[8] In our data, the number of training vacancies is highly correlated (0.83) with the number of beds, a proxy for hospital size. On the other hand, in a regression of the number of training vacancies per hospital on its lagged value, the number of beds, and the lagged proxy for hospital quality (the rank of the first candidate who chooses that hospital) shows that the latter is not significant. This suggests that the order in which physicians choose each hospital does not affect the number of training vacancies per hospital during our sample period.

[9] The standardized national exam is called the MIR exam after the Spanish acronym for "in-hospital resident physician."

[10] According to the Curso Intensivo Mir Asturias, a private academy that prepares candidates for the MIR exam, the excess of physicians over vacancies fluctuates between 19% and 78% of the vacancies during the last years (see http://www.curso-mir.com and tables: http://www.curso-mir.com/nuestros_result/01.htm and http://www.curso-mir.com/nuestros_result/02.htm).

[11] González (2004) claims that before 1996 some physicians chose from the available set any specialty in the hospital they wished to be located with the aim of requesting a change of specialty once the training period started. She also argues that regulation put in place in 1995 stopped these short-cuts by making it very costly to change specialties. We believe this strategic effect which, if at all, is only present in the 1995 data could only bias the hospital coefficients for those hospitals where switching specialties during training is perceived by the physician candidates to be easier.

4 Data

We use several data sets in our estimations. The main dataset contains the sequence of hospital-specialty choices made by physicians candidates from 1995-2000. These data are publicly available from the Spanish Ministry of Health and contain information on physician characteristics such as her position in the queue, gender, self-reported province of residence, college where she graduated, and her hospital-specialty choice.

As mentioned in the previous Section, physicians may drop out from the process and choose their outside option. We can identify the rank of those individuals who drop out from the process, however, there is neither information on their other characteristics nor on their outside option. A particular example of outside option is that of physicians who join compulsory military/social services and exercise their right to reserve a vacancy among those available to them. Because these physicians do not fill the vacancy immediately, physicians who follow in the queue will fill it.[12] This means that the set of vacancies available is not affected by these reservations, which is on average around 6.8% of the vacancies in any given year. As long as there are no informational cascades, choices made by reservationists do not affect the other physicians' choices. In which case, the absence of this information will not result in any bias.

After applying several filters to control for misspelling in the coding of the hospital or the specialty, our main dataset has around $2,585$ vacancies each year totalling $15,511$ observations from 1995 to 2000.[13] The total number of hospitals in the data set is 183. The number of hospitals varies by year because some hospitals are merged while others are split into different hospitals. We treat hospitals resulting from either merges or splits as new hospitals.[14]

Typically, for a particular hospital on average the number of vacancies per specialty per year is very small. Of all hospital-specialty combinations available, roughly 85% correspond to a single vacancy being offered and in 95% of the cases the number of vacancies is smaller than three, as figure 1 illustrates:

The average across hospitals of the number of vacancies (over all specialties) per year is 16.62 but this distribution is very disperse and skewed. For example, the median number of vacancies per hospital is only 7.17. The average number of specialties per hospital per year is on average relatively large (10.29) although the distribution is also relatively skewed with a median number of specialties per hospital of 7.00.

[12] Reservations from previous years that are filled during any year in our sample period are dropped from the dataset since those vacancies do not belong to the choice set of the other physicians during that year.

[13] After 1996 physicians who wished to become General Practitioners participated in a separate process which included a different exam and a different selection mechanism. This mechanism runs parallel to the other specialties. For this reason and because GP is a non-hospital based specialty we decided not to consider this specialty.

[14] The year with the lowest number of hospitals is 1995 with 146 while the year with the highest number is 2000 with 161.

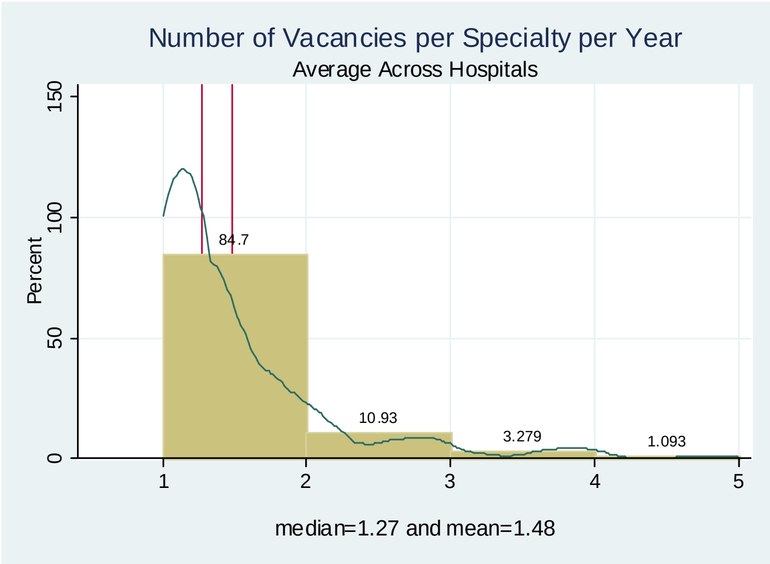


Figure 1:

There are large differences by specialty in the average number of vacancies as figure 2 shows:

In addition, there is a marked contrast between the distribution of male and female physician across specialties. Figure 3 shows the deviation of the gender ratio (i.e. the ratio of the share of female physicians to the share of male physicians) by specialty to the overall gender ratio in percentage terms. This index will be close to zero whenever the gender ratio in a given specialty is close to the overall gender ratio ($1.68 = 0.627/0.373$) as is the case for Internal Medicine. The most striking feature of Figure 3 is the large dispersion of female shares by specialty. For 35 out of 41 specialties, gender ratios deviate from the mean by more than 10%. At one extreme of the distribution are Allergy, Pediatrics, and Obstetrics & Gynecology where gender ratios exceed by more than 90% the overall ratio. At the other extreme, we find Urology, Neurological Surgery, Cardiovascular Surgery, and Orthopaedic Surgery where gender ratios are lower than the overall ratio by more than 75%.

From the information available in the physicians' data set it is possible to construct a variable which reflects a notion of geographical proximity between the physicians' residence and each of the hospitals. Whereas the hospitals' exact address is available, there are, however, two problems with the physicians' declared residence. First, physicians only declare the province of residence, not the city. Second, this variable may be measured with error as some physicians who have moved to attend college may declare

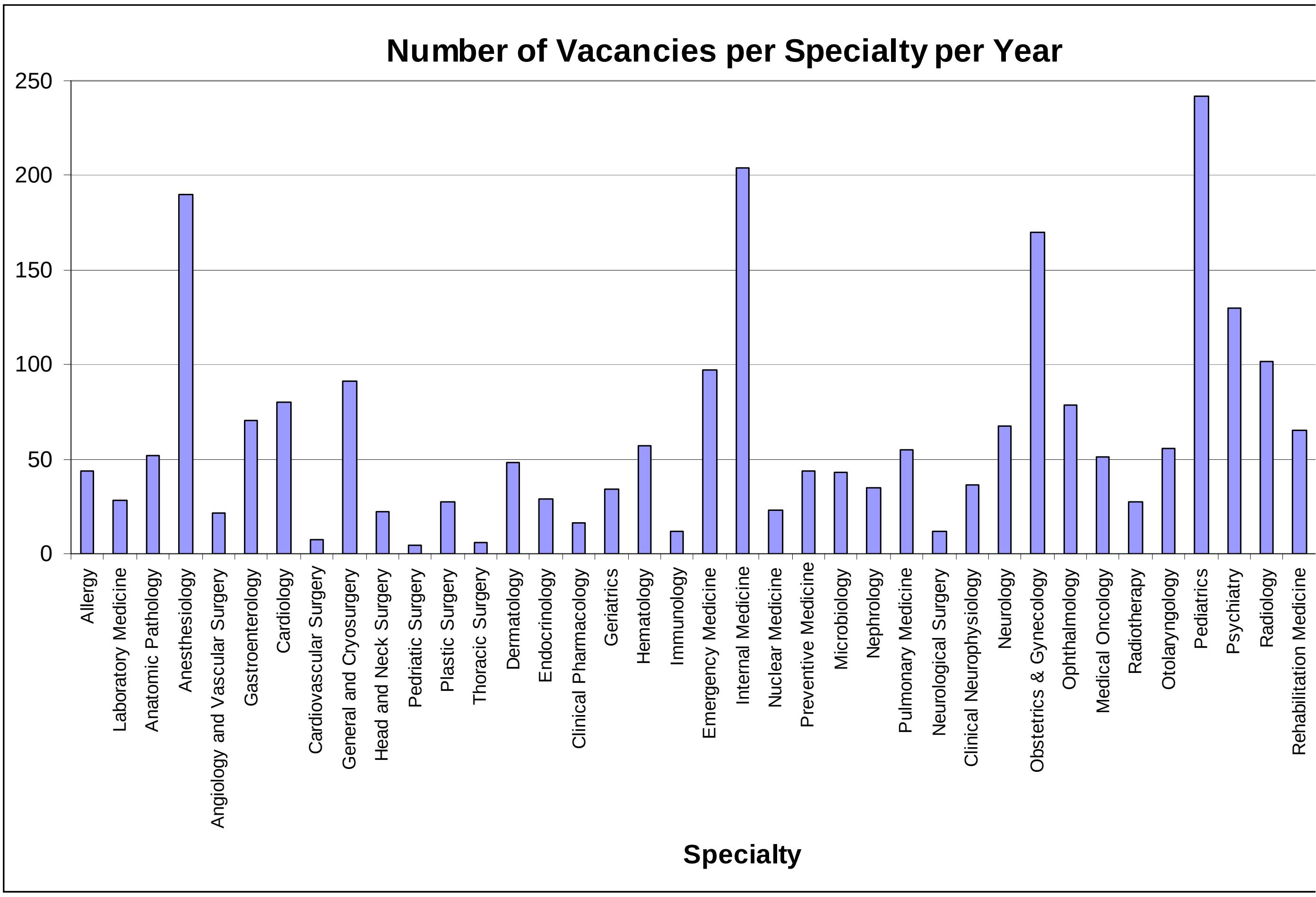

Number of Vacancies per Specialty per Year
250
200
150
100
50
0
Allergy
Laboratory Medicine
Anatomic Pathology
Anesthesiology
Angiology and Vascular Surgery
Gastroenterology
Cardiology
Cardiovascular Surgery
General and Cryosurgery
Head and Neck Surgery
Pediatric Surgery
Plastic Surgery
Thoracic Surgery
Dermatology
Endocrinology
Clinical Pharmacology
Geriatrics
Hematology
Immunology
Emergency Medicine
Internal Medicine
Nuclear Medicine
Preventive Medicine
Microbiology
Nephrology
Pulmonary Medicine
Neurological Surgery
Clinical Neurophysiology
Neurology
Obstetrics & Gynecology
Ophthalmology
Medical Oncology
Radiotherapy
Otolaryngology
Pediatrics
Psychiatry
Radiology
Rehabilitation Medicine
Specialty

Figure 2:

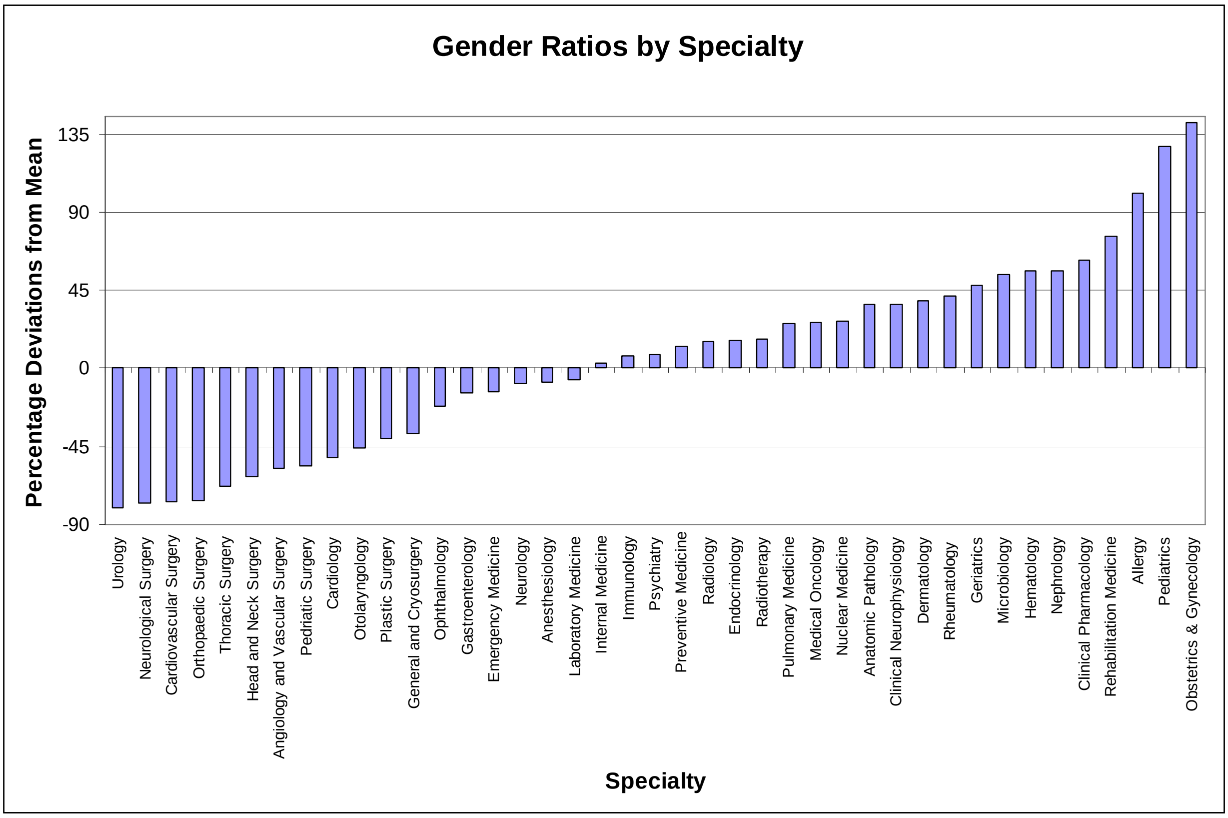

Gender Ratios by Specialty
Percentage Deviations from Mean
135
90
45
0
-45
-90
Urology
Neurological Surgery
Cardiovascular Surgery
Orthopaedic Surgery
Thoracic Surgery
Head and Neck Surgery
Angiology and Vascular Surgery
Pediatric Surgery
Cardiology
Otolaryngology
Plastic Surgery
General and Cryosurgery
Ophthalmology
Gastroenterology
Emergency Medicine
Neurology
Anesthesiology
Laboratory Medicine
Internal Medicine
Immunology
Psychiatry
Preventive Medicine
Radiology
Endocrinology
Radiotherapy
Pulmonary Medicine
Medical Oncology
Nuclear Medicine
Anatomic Pathology
Clinical Neurophysiology
Dermatology
Rheumatology
Geriatrics
Microbiology
Hematology
Nephrology
Clinical Pharmacology
Rehabilitation Medicine
Allergy
Pediatrics
Obstetrics & Gynecology
Specialty

Figure 3:

as province of residence either the college province or their parents' residence.[15]. This measurement error is potentially not independent from the physician's ability as studying in a good college enhances the likelihood of attaining a high score in the MIR exam. Consequently, good students, who aim to be trained at a top hospital, are more likely to move from their parents' home to study at a good college.

As an illustration of these issues consider the cases of Madrid and Barcelona, the two most populated provinces in Spain. Both Madrid and Barcelona act as focal points for medical students from other parts of the country. As can be seen from the statistics in Table **??**, the percentage of the overall college student population in Madrid over the total national (22.13%) is considerably higher than the percentage of citizens under 18 years of age over the total national (12.99%), which shows that Madrid's colleges attract students from the rest of the country. We partly observe this phenomenon in our sample, where 18.15% of physicians graduated in Madrid. From the census and the physicians' dataset information, we would expect the percentage of movers in Madrid to be between 30% and 70%. However, if we identify movers in our sample as those who declare province of residence different from the college's, then they are less than 10% of those who studied in Madrid. Thus, there seems to be serious underreporting of movers in Madrid. In contrast, the percentage of movers in Barcelona is within the expected interval, suggesting that there is no serious underreporting in Barcelona.

	% Movers[a,b]	% College Population		% Population under 18
		Sample[a] (medical schools)	Census 2001[c] (all colleges)	Census 2001[c]
Barcelona	14.56	11.73	14.18	10.94
Madrid	9.45	18.15	22.13	12.99

[a]Data from the physicians' data set.
[b]Movers are those whose province of residence differs from their college's.
[c]Data from the Spanish National Statistics Office (INE).

Table 1: Percentage of populations in Barcelona and Madrid over total national

To circumvent the problem of inconsistency and the potential biases associated with the residence location variable we use instead the location of the college where the physician studied. Near two thirds of physicians choose hospitals in their college location (58.11%) or in their declared province of residence (60.23%). This numbers increases to 68.94% and 72.10%, respectively, if we enlarge regions to larger administrative regions called Comunidades Autónomas[16]. If physicians decided upon location randomly then, for example, the percentage of physicians from Madrid, the region with the highest number of vacancies, who choose to stay in Madrid should be, on average, equal to the percentage number of vacancies in Madrid (14.5%). Yet, the figure for Madrid is 83.5%, well above the expected value for the random allocation. These figures suggest that hospital-in-same-region as college (residence) is an

[15]Data for Spain from the European Community Household Panel shows that around 80% of college students live in their parents' residence during our sample years .

[16]Spain is composed of 52 provinces. These are grouped into the 17 larger regions called Comunidades Autónomas.

important hospital characteristic affecting the physicians' choice. The preference to stay in the same region will be introduced in our model in the form of "nests" a la McFadden. We employ the city-level travel-time-to-work dataset used in Holl (2004) to define the nests.

We complement the main dataset with data from the Spanish National Catalogue of Hospitals (*Catálogo Nacional de Hospitales*), which is published annually by the Ministry of Health. These data contain hospital characteristics such as location (city), number of beds and quantity of available technical equipment (e.g. in our sample: the availability ranges from most available equipment, emission tomography equipment, for which 88.52% of hospitals have at least one, to least available equipment, extracorporeal shock wave lithotripsy, for which only 28.42% of hospitals have at least one). Finally, we also gathered provintial level data on annual average housing prices, the number of physicians per 1000 inhabitants, and unemployment rates from the Spanish National Statistics Office (INE) to control for regional differences in living costs.

5 The Econometric Framework

In this section we propose the framework to construct a ranking of hospitals based on individual choices made by recently graduated physicians. First, in Subsection 5.1 we simplify the decision model by assuming that geographical proximity to hospital plays no role in the physicians decision and that physicians must choose one of the available vacancies. Subsequently, in Subsection 5.2, we extend the model to incorporate the preference to stay in the same region.

5.1 The Basic Model

Physicians are ordered according to their average score. Let $i = 1, ..., N$ simultaneously identify a physician as well as her position in the rank. For example, $i = 1$ ($i = N$) denotes the physician who obtained the highest (lowest) score. Each physician i chooses one hospital-specialty combination from those available to her. Hospitals may open more than one vacancy for a given specialty. Denote by C_i the set of all hospital-specialty combinations available to physician i. If i chooses a hospital-specialty combination for which there are at least two remaining vacancies then $i + 1's$ choice set is identical to C_i. Eventually, however, as hospital-specialty vacancies are exhausted, choice sets must shrink: C_1 contains all possible hospital-specialty vacancies and C_N has a single vacancy available. Hence, choice sets satisfy the following:

$$C_i \supseteq C_{i'} \text{ for } i' > i. \tag{1}$$

We model physician preferences according to the stochastic random utility model where U_{ij} represents the utility to physician i from selecting hospital-specialty combination j. U_{ij} is decomposed into a deterministic component V_{ij} and a iid stochastic term ε_{ij} :

$$U_{ij} = V_{ij} + \varepsilon_{ij} \tag{2}$$

where ε_{ij} is assumed to follow an extreme value distribution, i.e. $\Pr(\varepsilon_{ij} \leq x) = \exp(-\exp(-x))$. McFadden (1974) shows that the probability that physician i chooses the hospital-specialty combination $j^* \in C_i$ is given by:

$$P_{ij^*} = \frac{\exp(V_{ij^*})}{\sum\limits_{j \in C_i} \exp(V_{ij})}, \text{ for } j^* = 1, ..., J_i \tag{3}$$

where J_i denotes the number of elements in C_i.

Let $\pi = (\pi_1, ...\pi_N)$ be the observed sequence of physician choices where π_i indexes the hospital-specialty combination chosen by physician i. The assumption on iid error terms implies that the probability of observing a given sequence of hospital-specialty choices, $\Pr(\pi)$, is a sequence of independent multinomial logit models:

$$\begin{aligned} \Pr(\pi) &= \Pr(U_{1\pi_1} \geq U_{1j} | \forall j \in C_1) \times \Pr(U_{2,\pi_2} \geq U_{2j} | \forall j \in C_2) \\ \times ... \times \Pr(U_{N-1,\pi_{N-1}} &\geq U_{N-1j} | \forall j \in C_{N-1}) \end{aligned} \tag{4}$$

Equation (4) is formally similar to the "exploding conditional multinomial logit" of Chapman and Staelin (1982) and Bradlow and Fader (2001). Our model, however, differs because the ranking π is obtained from the choice of different individuals and each physician chooses one option only and does not provide a complete ranking of hospital-specialties combinations.[17] Since physicians choose their preferred hospital-specialty combination from their choice set C_i, the sequence of observed hospital-specialties choices π provides useful information to construct a ranking of hospital-specialties combinations.

Physician preferences over hospital-specialty combinations depend on interactions of hospital-specialty characteristics and individual characteristics. Let j be an index of the combination of hospital h and specialty s.We assume that V_{ij} depends linearly on a vector x_h of hospital characteristics, a vector x_s of specialty characteristics, and finally on a vector x_{ij} of interactive variables relating physician i to hospital-specialty j:

$$U_{ij} = x_h \beta_h + x_s \beta_s + x_{ij} \beta_j + \varepsilon_{ij}. \tag{5}$$

In order to construct an appropriate hospital ranking we need to infer from the physicians' choices a measure of the hospitals' quality as perceived by the physicians. Estimates of hospital dummy coeffi-

[17] In constrast, Chapman and Staelin's (1982) model assumes that each individual provides a ranking of choices. Bradlow and Fader (2001) data, on the other hand, consists of weekly observations of the pop songs hit list Billboard "Hot 100". Their ranking is provided by the level of record sales and, therefore, implicitly aggregates the choice of many different individuals. They do not model the aggregation of preferences but instead work as if a single entity gives the ranking.

cients provide a reasonable aproximation for a measure of the latter.[18] Nevertheless, four shortcomings should be acknowledged: first, hospital quality may vary across specialties. This could be addressed by including hospital-specialty dummy variables. In our data set, for some specialties and hospitals, however, there are not enough observations to obtain precise estimates (see figure 1). Second, as commented in the Introduction, hospitals' quality as perceived by the physicians may differ from hospitals' quality from the point of view of society. For example, physicians may prefer to be placed in larger hospitals because they may learn more. Hence, if we do not control for size, the estimated hospital dummy coefficients for larger hospitals would be an upward-biased estimate of quality in the latter sense. In order to minimize this potential bias, the Results Section will also present estimates controlling for hospital observable characteristics such as size and the availability of sophisticated equipment. Third, controlling for time-variant hospital characteristics does not capture unobservable time invariant city characteristics that contribute to the attractiveness of hospitals in the city. In this case, hospital dummy coefficients would not only reflect relative hospital training quality but also embody physicians' valuations for these city characteristics. In the Results Section we present city-level hospital rankings which are immune to this shortcoming. Fourth, it is conceivable that information about hospital quality may depend on some characteristics of the physician, such as her ranking. In that case, we should control for individual heterogeneity in preferences, for example, allowing for individual specific hospital coefficients, β_{ih}. However, given the large number of available vacancies in our data set, we decided, for computational reasons, to model heterogeneity as a function of the physician's ranking. In particular, we allow for different hospital dummy coefficients for the physicians in the top quartile of the ranking.

The Likelihood function for the basic model takes the form:

$$L(\pi|\beta) = \Pr(\pi) = \prod_{i=1}^{N} \frac{\exp(V_{ij^*}(\beta))}{\sum\limits_{j\in C_i} \exp(V_{ij}(\beta))} \tag{6}$$

where $\beta = (\beta_h', \beta_s', \beta_j')'$ is the vector of parameters. An important drawback from this model is that it assumes the *Independence of Irrelevant Alternatives* (IIA) holds. The way we introduce preferences for location in the next subsection partially relaxes this assumption.

5.2 Introducing Location

When physicians choose a particular hospital they are implicitly choosing their residence for the following four to five years. As already stated, physicians have a preference to stay in the province where they graduated or reside. Table 2 shows the percentage of physicians who choose a vacancy in the same province as their college or declared residence for different rank quartiles. Table 2's statistics may struck as counter-intuitive since one expects the highest ranked physicians to give a higher weight to

[18] Note that the incidental parameter problem present in fixed effects estimators within non-linear models does not arise in this context as the number of hospitals does not increase with the number of observations within a year (i.e. with the number of vacancies).

hospital quality and, therefore, to be more willing to move to a different region in order to be trained at a better hospital. However, since college quality is not evenly distributed across the country, as Figure 4 suggests[19], top students have already moved to study medicine in the best colleges in order to improve their position in the MIR ranking.[20] Thus, if best hospitals are located in the same regions as the best colleges, top ranked physicians do not need to move again. This explanation highlights that the variable hospital-in-same-region-as-college is likely to be endogenous since unobserved factors that affect hospital choice also affect college choice. Hence, the introduction of hospital-in-same-region-as-college in the physicians' utility, as in Avery et al. (2005), would lead to biased estimates. The nested-logit framework presented below allows us to introduce preference for proximity without causing a bias in the estimated coefficients.

	Rank quantiles			
	25%	25% – 50%	50% – 75%	75% – 100%
hospital in same province as college	67.63	63.38	54.31	47.11
hospital in same province as residence	67.37	65.01	57.81	50.72

Table 2: Percentage of Physicians in each quantile that choose a hospital in the same region as college or residence

We introduce location by extending the multinomial logit framework to a nested logit (McFadden, 1978) where the physician's hospital-specialty choice also involves a location choice. Although the choice of location and hospital-specialty is simultaneous, the nested logit can be interpreted as a sequential choice model where one first chooses location (or *nest*) and then a hospital-specialty within that location.[21]

Nests based on geographical proximity variables play an important role in the identification of hospital quality because part of the identification comes precisely from those physicians who choose a hospital with a location far away from that of the college. Physicians who are highly ranked will have more options and, therefore, a movement to another location reveals a stronger preference for the chosen hospital. On the contrary, physicians who are lower ranked may have to move to another location because there are fewer options to choose from. Therefore, movements made by highly ranked physicians help identify relative hospital quality.

[19] Figure 4 shows the percentage of physicians ranked amongst the top 25% among those who choose a vacancy by region of college. Regions with only one college are left out of the figure to comply with a confidentiality agreement. Nonetheless, the range of values for the excluded regions (20.66% versus 31.53%) is similar to the range shown. The physicians from the "best performing" region fare roughly 50% better than those coming from the "worst performing" region.

[20] For example, as shown in Table 1, Madrid and Barcelona attract a large number of students from other regions.

[21] Conceivably we could extend this framework to the hospital-specialty choice with a three-level nested logit where locations and specialties define nests. This would be a more general model than the one we propose here. However, given the large number of specialties, the number of parameters would make the estimation computationally burdensome.

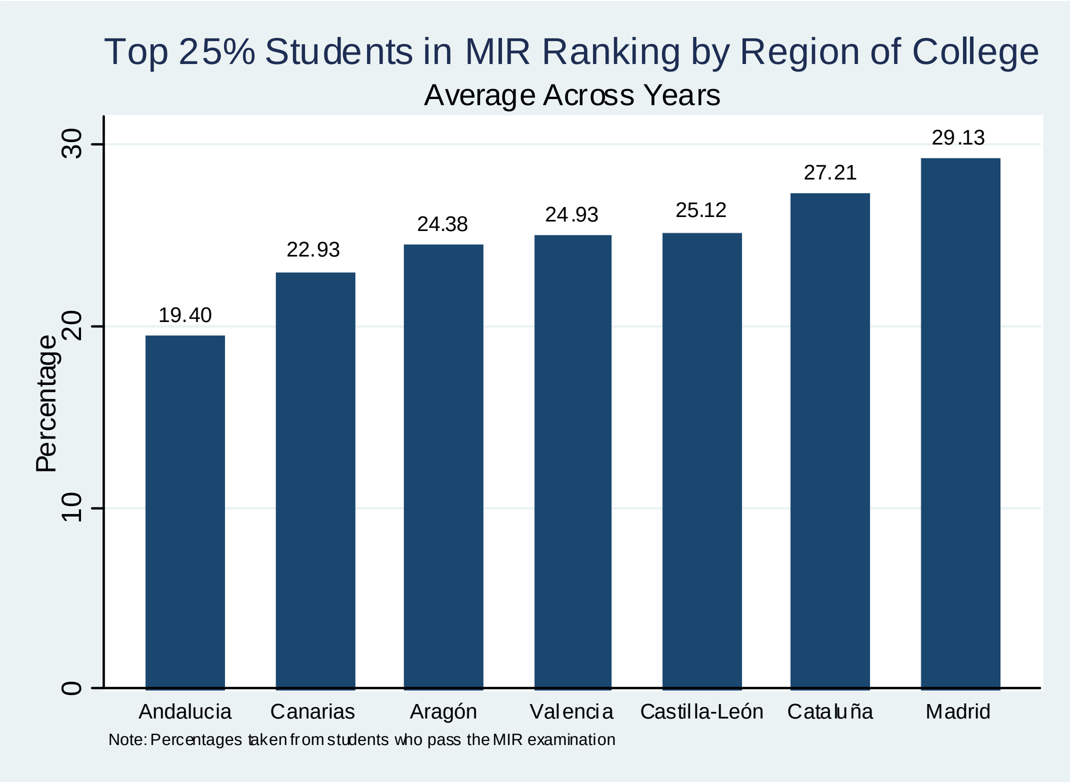

Top 25% Students in MIR Ranking by Region of College
Average Across Years
Percentage
0
10
20
30
19.40
22.93
24.38
24.93
25.12
27.21
29.13
Andalucia
Canarias
Aragón
Valencia
Castilla-León
Cataluña
Madrid
Note: Percentages taken from students who pass the MIR examination

Figure 4:

Relative to the standard multinomial logit, the nested logit has the additional advantage of relaxing partially the IIA. It only requires the IIA to hold for alternatives within the same nest.

In the simplest specification, we partition the set of all hospital-specialty combinations available for each physician i into two nests, B_{ik}, $k = 1, 2$. Without loss of generality, nest 1 includes hospitals located close to the physician's location (college) while nest 2 includes hospitals located elsewhere. Thus, nests are defined as an interaction between hospital and physician locations. Using the city-level data on travel-time-to-work, a hospital is classified into nest 1 whenever it is located within 45 minutes of college.[22]

As in the previous Section, choice sets, C_i, are individual specific and shrink as hospital-specialty combinations are exhausted along the process. The physician's utility from choosing alternative j belonging to nest k, $j \in B_{ik} \subset C_i$, is:

$$U_{ij} = V_{ij} + \varepsilon_{ij} \tag{7}$$

where V_{ij} depends linearly on a vector x_h of hospital characteristics, a vector x_s of specialty characteristics, and finally on a vector x_{ij} of interactive variables relating physician i to hospital-specialty j just as in equation (5). The difference with the multinomial model lies in the distribution of the error term, which now depends on the definition of nests. To be precise, the vector $\varepsilon_i = \left(\varepsilon_{i1}, ...\varepsilon_{ij}, ...\varepsilon_{iJ(i)}\right)$ has cumulative distribution:

$$\Pr(\varepsilon_i \leq a|C_i) = \exp\left(-\sum_{k=1}^{2}\left(\sum_{j \in B_{ik}}^{2} e^{-\frac{a_j}{\lambda_k}}\right)^{\lambda_k}\right). \tag{8}$$

Importantly, the nested-logit model allows us to introduce the preference for proximity without causing a bias in the estimated coefficients. To see this, note that the proximity variable, which defines individual-specific nests, does not affect the deterministic component of utility directly, i.e. is not included in V_{ij}, and only affects the distribution of the error term ε_i.[23]

If $\lambda_k = 1$ for all k then the elements of ε_i are independently distributed and the model coincides with the multinomial logit. [24] It can be shown (McFadden, 1974) that the probability that physician i

[22] Initial versions of this paper considered using administrative regions -so called *Comunidades Autonomas*- to define nests. The classification of hospitals into nest 1 and nest 2 remains the same for 89% of the observations relative to the travel-time-to-work regions. The fact that some cities are near province boundaries and people move acrosss them, suggests that the travel-time-based nests are more realistic. In addition, while point estimates are similar under the two specifications, standard errors were larger under the administrative-based nests.

[23] Avery et al. (2005) introduce the same-region-as-residence variable as a component of V_{ij} in their model of college choice. This specification causes them no bias since, in their data, the location of residence is precisely measured and exogenous.

[24] Values of $\lambda_k < 0$ are not consistent with utility maximization. Train (2003) asserts that values of $\lambda_k \in (0, 1]$ are always consistent with utility maximizing behavior while $\lambda_k > 1$ is consistent for a range of values of the independent variables $X's$.

chooses the hospital-specialty combination $j^* \in B_{ik} \subset C_i$ is given by:

$$P_{ij^*} = \frac{e^{\frac{V_{ij^*}}{\lambda_k}} \left(\sum_{j \in B_k} e^{\frac{V_{ij}}{\lambda_k}} \right)^{\lambda_k - 1}}{\sum_{l=1}^{2} \left(\sum_{j \in B_l} e^{\frac{V_{ij}}{\lambda_l}} \right)^{\lambda_l}}, \text{ for } j^* \in B_{ik} \tag{9}$$

The parameter λ_k is a measure of the degree of independence among the elements of ε_i in nest k (see for example Train, 2003, pp 83). Greater dependence between options in nest k (lower λ_k) increases the spread of odds ratios within that nest. Intuitively, the higher the correlation between $\varepsilon_{ij}'s$, the more likely the same hospitals are "winners" or "losers" for different physicians.

The difference $\lambda_1 - \lambda_2$ measures the relative preference for nest 1. This can be seen in an extreme example. Suppose both nests have the same number of vacancies, N, and $V_{ij} = V$ for all i and j. Then the probability of choosing any vacancy in nest 1 relative to the probability of choosing any vacancy in nest 2 equals $N^{(\lambda_1 - \lambda_2)}$. Moreover, a value of λ_1 close to 1 implies that alternatives in nest 1 are seen as more substitutable, in the sense that it lowers the spread of the odds ratio, because proximity to hospitals is a crucial factor in the physicians' choice. For a given value of λ_2, a higher value of λ_1 increases the probability of choosing any hospital in nest 1 vis-a-vis any hospital in nest 2, and simultaneously decreases the difference in odds ratios between hospitals in nest 1. Consequently, we expect $\lambda_1 - \lambda_2 > 0$, and λ_1 to be around 1 in our application. Our estimates of $\lambda_1$and λ_2 follow this pattern.

We extend the simplest specification in several ways. First, we add a third nest to account for potential language barriers within the national system.[25] Nest 1 still refers to all hospitals within $45'$ from college. Nest 2 now includes all hospitals beyond $45'$ which are located in a province with a set of official languages included in the set of official languages in the province of the physician's college. Nest 3 includes all other hospitals. Consider the following example with only three cities which are located more than $45'$apart: Barcelona, Bilbao, and Madrid. In Barcelona there are two official languages: Catalan and Spanish (Castilian). In Bilbao there are also two official languages, but they are Basque and Spanish. Finally, in Madrid there is only Spanish. Thus, a student from Madrid would have all available vacancies from hospitals at Bilbao and Barcelona in nest 3. In contrast, a student from Bilbao would have in nest 3 only hospitals from Barcelona and a student from Barcelona would have in nest 3 only hospitals from Bilbao. In the sample, 60.59% of physicians choose a vacancy from nest 1, 31.26% from nest 2, and 7.76% from nest 3.

[25] Many physicians who finish the training period to become specialist apply to permanent positions within the public health system. Regional authorities in multilanguage regions often include proficiency levels in local languages as merits in the competition for those positions. This has allegedly crowded-out candidates in some regions (see, for example, http://www.elpais.com/articulo/sociedad/exigencia/euskera/agrava/falta/medicos/Pais/Vasco/elpepusoc/20071102elpepisoc_5/T

6 Results

Table 3 gives an overview of all model specifications considered and Table 4 summarizes the estimation results. Model specifications can be classified into two groups. In the first group, Models 1 and 2 in Table 3, we only allow for two nests. Nest 1 includes all hospitals located within 45 minutes from the physician's college location while nest 2 includes all other hospitals. In the second group, Models 3, 4, and 5, we split nest 2 into two nests as explained in the previous section. We redefined nest 2 to include all hospitals beyond the 45−minute commuting threshold but within the language boundaries. In the new nest 3 we include those hospitals which are both beyond 45 minutes and outside the language boundaries. We also consider different variable specifications with respect to vector x_h in (5). In the simplest specification, Models 1 and 3, x_h contains only hospital dummies. The rest of the models, Models 2, 4, and 5, expand the set of controls to include hospital characteristics such as number of beds (in thousands) and dummies of technical equipment, and also regional characteristics such as housing prices, unemployment rates, and the number of physicians per thousand inhabitants all measured at the provincial level. All models, contains specialty dummies in x_s as in (5) and an interaction term between physician's gender and specialty in x_{ij}. The motivation for the latter is taken from the pattern of segregation between female and male specialty choices described in the Data Section. Finally, physicians may have heterogenous preferences on hospitals leading to different rankings. Among the potential factors affecting these differences, the rank position of the physician is an obvious one since top ranked physicians have a stronger incentive to gather information at the same cost.[26] We incorporate this effect in the model in a crude but operational way by allowing different hospital dummy coefficients between physicians in the top 20% and all the other physicians in Model 5.[27]

Description	Model 1	Model 2	Model 3	Model 4	Model 5
hospital dummies (ref. H13)	yes	yes	yes	yes	yes
specialties dummies (ref. O&G)	yes	yes	yes	yes	yes
specialties ×woman (ref. O&G)	yes	yes	yes	yes	yes
hospital and regional characteristics	no	yes	no	yes	yes
hospital ×rank dummy (top 20%)	no	no	no	no	yes
nest 1	$\leq 45'$ from college		$\leq 45'$ from college		
nest 2	$> 45'$ from college		$> 45'$ from college&within language boundaries		
nest 3	–	–	$> 45'$ from college&outside language boundaries		

[26] In informal conversations with hospital managers, we were told that every year, the best candidates are invited to visit the hospital's departments.

[27] Some hospitals are never part of the choice set for the low rank physicians because all vacancies have been taken by top physicians. There is also a set of hospitals which are never chosen by top physicians. In none of the these cases, it is possible to separately identify hospital dummies for top and low rank physicians. For these hospitals, we impose the same hospital dummy coefficient for all physicians. Overall, for 112 hospitals identification of the hospital dummy coefficient for top physicians is possible. Hereafter, the estimates of the 112 hospitals for the bottom 80% physicians together with the estimates for the 61 hospitals for which no interaction is possible, will be reffered to as the bottom 80% estimates.

Table 3: Description of Models

Table 4 summarizes the estimation results. For brevity, we do not show the estimated coefficients for all hospital dummies and specialties. Because the highest and lowest hospital estimated coefficients correspond to the same two hospitals regardless of specification, we also show them in Table 4. For Model 5, we only show the estimated hospital dummy coefficients for the best and the worst hospital associated with the bottom 80% physicians. None of the estimated coefficients for the interaction between top physicians and hospital dummy coefficients (i.e., the differences in valuations between top 20% and bottom 80% physicians) are statistically significant. Moreover, results for the rest of the variables are similar to Model 4. Thus, we find no evidence of heterogenous preferences between the best physicians and all the others.

Hereafter, we focus our discussion on models 1 to 4. The inclusion of controls widens the gap between the best and the worst hospital, especially in terms of odds ratios. The inclusion of a third nest widens this gap even further. Concerning the results for the set of controls, the coefficient estimates for the regional unemployment rates, number of beds, and the presence of extracorporeal shock wave lithotripsy equipment are the only ones statistically significantly different from zero. These results are not surprising as coefficient dummies for equipment availablity only embody the impact of changes in equipment presence over the sample on the probability of choosing a hospital. Given that these variables change over the sample in between 21 and 43 percent of all hospitals, and correlations among these controls range between 0.18 and 0.65, we should not expect lack of identification of these effects. Note that these results do not imply that equipment availability is unimportant as the impact of average equipment is already embodied in the hospital dummies.

The results on the estimates of the nests parameters shown in Table 4 deserve several comments. First, they are very precisely estimated. Second, they show physicians value geographical and language proximity ($\widehat{\lambda}_1 > \widehat{\lambda}_2 > \widehat{\lambda}_3$). Third, estimates for $\widehat{\lambda}_1$ are very close to one (statistically different from 1), indicating that hospitals in nest 1 are perceived by physicians as more substitutable. Although estimates for $\widehat{\lambda}_1$ exceed 1 statistically significantly when splitting hospitals into three nests (Model 3, 4, and 5), estimates are still consistent with utility maximizing behavior (see Herriges and Kling, 1996).

description	Model 1	Model 2	Model 3	Model 4	Model 5
best hospital	2.5175*** (0.2042)	2.5723*** (0.2138)	3.200*** (0.2620)	3.2440*** (0.2710)	3.2881*** (1.1702)
worst hospital	−1.5018*** (0.2555)	−1.5223*** (0.2625)	−1.9166*** (0.3110)	−1.858*** (0.316)	−1.7993*** (0.3108)
number of beds (in thousands)	–	0.0792 (0.1028)	–	0.2141* (0.1304)	0.1722 (0.1292)
tomography	–	0.0288 (0.0257)	–	0.0263 (0.0334)	0.0174 (0.0332)
magnetic ressonance	–	0.0229 (0.0168)	–	0.0304 (0.0221)	0.0320 (0.0220)
gamma camera	–	0.0028 (0.0278)	–	−0.0007 (0.0359)	−0.0049 (0.0357)
hemodynamic monitoring equipment	–	−0.0229 (0.0234)	–	−0.0297 (0.0308)	−0.026 (0.0306)
angiography digital subtraction	–	−0.0216 (0.0270)	–	−0.0246 (0.0338)	−0.0244 (0.0338)
extracorporeal shock wave lithotripsy	–	0.0452** (0.0235)	–	0.0508* (0.031)	0.0482 (0.0308)
cobalt treatment equipment	–	−0.0278 (0.0211)	–	−0.0253 (0.0271)	−0.0214 (0.0268)
particle accelerator	–	0.0070 (0.0203)	–	0.0061 (0.0262)	0.0038 (0.026)
index of housing prices (1990=0.1)	–	0.0504 (0.5223)	–	0.0892 (0.6476)	0.0598 (0.6457)
physicians per 1000 inhabitants	–	−0.0184 (0.0249)	–	0.0032 (0.032)	−0.0059 (0.0319)
unemployment rate	–	−0.0097*** (0.0037)	–	−0.0195*** (0.0048)	−0.0179*** (0.0048)
λ_1	0.9406*** (0.0107)	0.9431*** (0.0107)	1.1085*** (0.0089)	1.1061*** (0.0089)	1.1007*** (0.0092)
λ_2	0.6212*** (0.0076)	0.6229*** (0.0076)	0.7486*** (0.0070)	0.7471*** (0.0071)	0.7416*** (0.0073)
λ_3	–	–	0.6182*** (0.0079)	0.6154*** (0.0079)	0.6093*** (0.0081)
Nobs	15511	15511	15511	15511	15511
Wald test for equality of all λ's	5954.5	6002.4	15761.9	15686.4	14910.9
Wald test for $\lambda_2 = \lambda_3$	–	–	862.4	878.0	847.0
Log-Lik.	−73993.38	−73976.59	−74250.92	−74232.74	−73986.93

Table 4: Some coefficient estimates

Figure 5 shows specialty dummy coefficient estimates (relative to obstetrics-gynecology) by gender for models 2 and 4 (the coefficient estimates for models 1 and 3 are graphically identical to the results for models 2 and 4, respectively).[28] Specialties are sorted in descending order according to the estimates for males in model 2. Thus, cardiovascular surgery is the preferred specialty for males while microbiology is the least preferred of all specialties. Overall, male physicians value fifteen specialties significantly above obstetrics-gynecology (in the graph, all specialties to the left of pedriatic surgery) and twenty-three specialties significantly below obstetrics-gynecology (all specialties to the right of otolaryngology). Except for pedriatics, female physicians value (relative to obstetrics-gynecology) all specialties statistically less than male physicians. Further, gender coefficient estimates are correlated on the right side of the graph while this pattern does not occur on the left side of the graph. Thus, those specialty most valued (relative to obstetrics-gynecology) by male physicians are not systematically most valued by female physicians. In particular, only seven out of the fifteen specialties more valued than obstetrics-gynecology by male physicians are also significantly more valued by female physicians.[29]

The continuous line in figure 5, labelled "Gender Ratios" depict the standardized values shown in figure 3. These values show the deviation of the gender ratio (i.e. the ratio of female to male physicians) by specialty to the overall gender ratio in percentage terms, which results from an interaction of tastes and availability. We interpret the estimated specialty dummy coefficients as showing only gender-specific preferences. This graph allows us to understand the apparent seggregation shown in figure 3. For example, urology's gender ratio is the lowest in our sample while it is neither the most valued by male nor the least valued by female physicians. Its very low gender ratio is nevertheless explained by the relative valuation of urology across gender, i.e. it ranks 12*th* among male physicians versus 26*th* among female physicians. At the opposite end of gender ratios lies pedriatics. Its gender ratio is after obstetrics-gynecology the highest in our sample because it ranks 14*th* among male physicians versus 6*th* among female physicians. In order to assess to what extent physicians are able to choose vacancies according to their preferences, we correlate the observed gender ratios and the coefficient estimates for the interaction between specialties and gender. In all model specifications the correlations are higher than 0.94, suggesting that the ranking of physicians precludes neither male nor female physicians from choosing according to their preferences.[30]

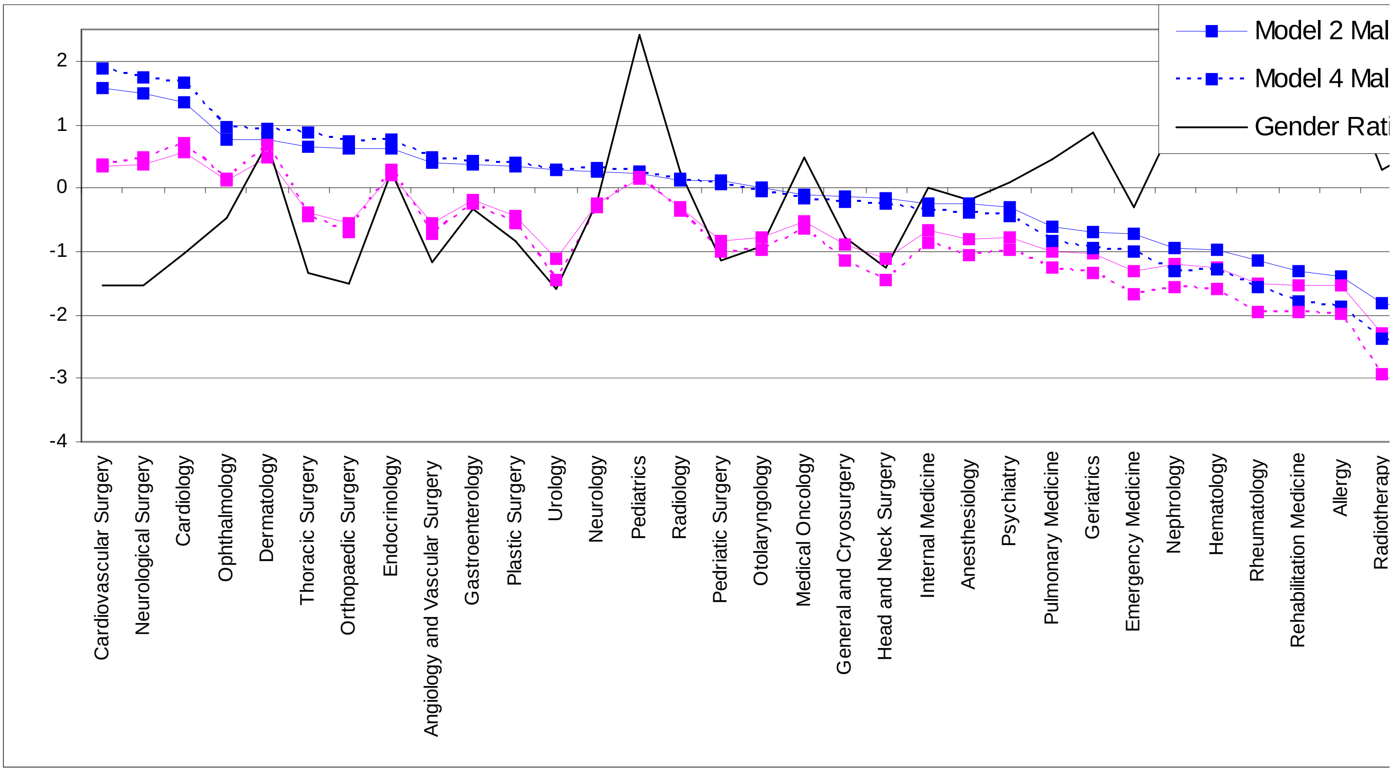
6.1 Results on Hospital Dummy Coefficients and Rankings

Figure 6 shows the estimated hospital dummy coefficients and their 90 percent confidence intervals for the first four specifications. The order of the hospital coefficients drawn in all graphs follows the ranking order from Model 1. We interpret these coefficients as quality differentials relative to the reference hospital (for which quality has been normalized to zero). Although these coefficients vary

[28] For clarity, we do not show standard deviations in figure 5. Results are available upon request.

[29] These can be easily identified in the graph as the ones for which the estimates are positive, with the exception of Ophtalmology for Model 4, which is only significant at the 15%.

[30] If, say, female physicians were systematically lower ranked relative to male physicians and could only choose the unwanted specialties, then we would not be able to identify the coefficient on the interaction of specialty and gender. Indeed, the data shows that there is a substantial overlap by gender in the physician rank. For example, amongst the top 100 physicians, 56% are female.



Model 2 Mal
Model 4 Mal
Gender Rati
2
1
0
-1
-2
-3
-4
Cardiovascular Surgery
Neurological Surgery
Cardiology
Ophthalmology
Dermatology
Thoracic Surgery
Orthopaedic Surgery
Endocrinology
Angiology and Vascular Surgery
Gastroenterology
Plastic Surgery
Urology
Neurology
Pediatrics
Radiology
Pedriatic Surgery
Otolaryngology
Medical Oncology
General and Cryosurgery
Head and Neck Surgery
Internal Medicine
Anesthesiology
Psychiatry
Pulmonary Medicine
Geriatrics
Emergency Medicine
Nephrology
Hematology
Rheumatology
Rehabilitation Medicine
Allergy
Radiotherapy

Figure 5:

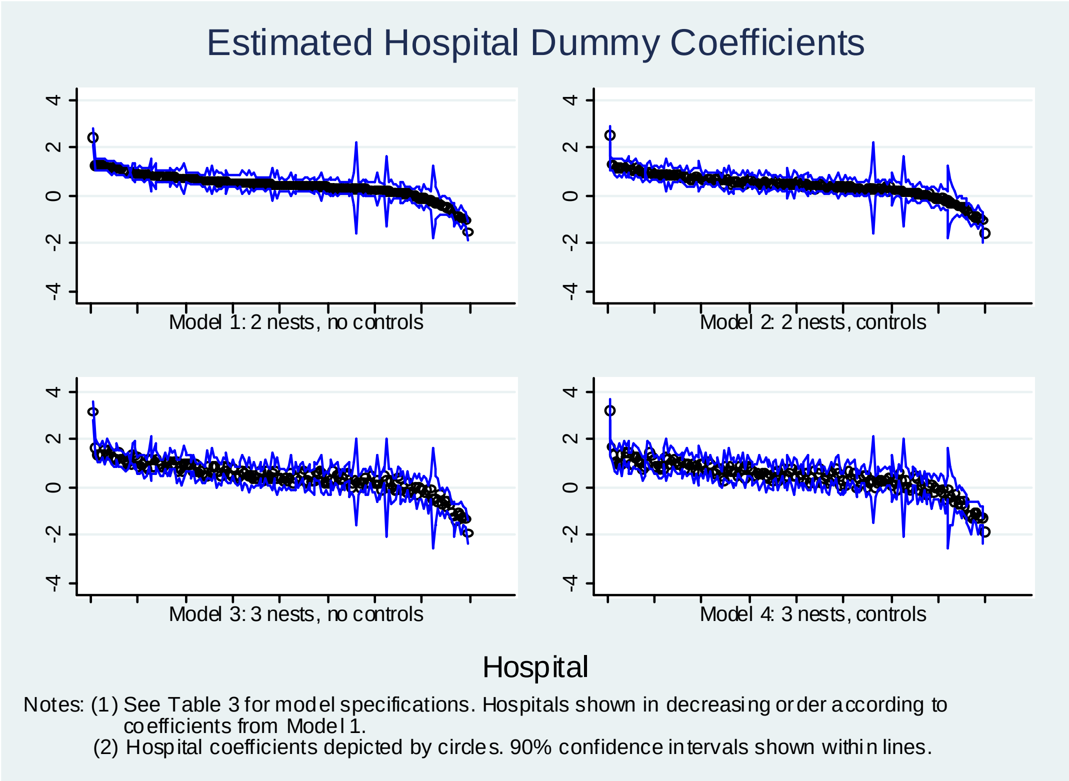


Figure 6:

accross specifications, the overall ranking of hospitals is fairly stable as can be seen by the negative slope in all the graphs in Figure 6. It is also apparent from the graphs that the inclusion of controls affects the ranking by less than the inclusion of a third nest as the slope becomes more irregular. The probability of choosing the best hospital relative to the worst hospital ranges from 55.7 (Model 1) to 166.8 (Model 4), and these odds ratios are statistically significantly different from zero for all specifications.

We further check the similarity of the hospital rankings accross all specifications by computing two alternative correlation measures between estimates of the hospital dummy coefficients. We show the results in Table 8. In the upper panel of Table 8 we report Kendall's τ rank correlation, which compares rankings rather than point estimates. The second panel shows simple correlations. "Sample" refers to the negative of the average across years of the order of the first physician to choose a given hospital (any specialty). Simple and rank correlations between all model specifications are very high and significantly different from zero. The lowest correlations always involve "Sample" and they are decreasing with the complexity of the other model. These results suggest that using "Sample" as an indicator of hospital quality would result in a rather different ranking than using any of the other estimated rankings. Simple correlations between the estimated coefficients of the different model specifications are always above 0.95. Kendall's τ rank correlations follow a similar pattern but show wider quantitative differences among them. A value of 0.754 means that around 88% of pairwise ranking comparisons are concordant. In contrasts values of 0.451 and 0.923 imply percentages of concordant pairs of 73% and 96% respectively. These results suggest that adding complexity to the model implies some differences in the rankings obtained, but overall, rankings are robust across different model specifications.[31]

	Model 1	Model 2	Model 3	Model 4	Model 5
	Kendall's τ rank correlation				
Sample	0.493	0.477	0.477	0.451	0.459
Model 1		0.923	0.786	0.754	0.758
Model 2			0.756	0.771	0.775
Model 3				0.883	0.880
Model 4					0.960
	Simple correlation				
Sample	0.572	0.563	0.565	0.543	0.552
Model 1		0.995	0.965	0.952	0.953
Model 2			0.959	0.960	0.962
Model 3				0.987	0.985
Model 4					0.998
Notes: (a) p-values for the test of independence between					
the rankings are always smaller than 0.0001. These are not					
corrected for the fact that rankings are estimates.					
(b) Model 5 refers to rankings obtained from hospital dummy					
coefficients for the bottom 80 percent physicians.					

Table 5: Kendall's tau rank correlation between hospital rankings

It could be argued that controlling for relative location using nests and time variant hospital characteristics does not capture unobservable time invariant city characteristics that contribute to the attractiveness of hospitals in the city. In this case, the interpretation of hospital dummy coefficients differentials bewteen hospitals located in different cities may confound hospital quality and city attractiveness. Still, within-city rankings should reflect only hospital differentials regarding physicians' preferences. In our dataset, there are 32 out of 67 cities with more than one hospital, nine with more than three hospitals, and only three cities with more than seven (Barcelona, Madrid, and Valencia). We can assess for the 32 cities whether hospital coefficient estimates are significantly different with

[31] For the 112 hospital for which we allow for a different valuation for the top 20% physicians, the rank correlation and the sample correlation between the rankings obtained from the estimates for the top 20% and the bottom 80% physicians are 0.735 and 0.904, respectively. This shows that not only are the interaction terms statistically insignificant, but they are small.

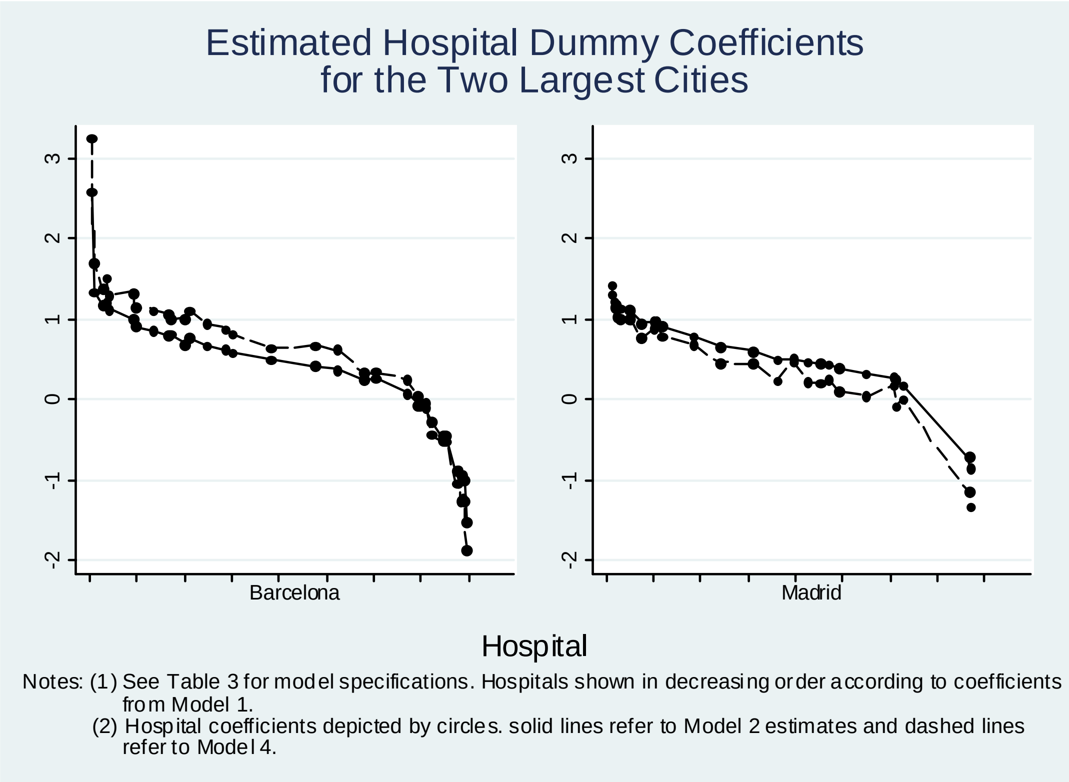


Figure 7:

respect to city-specific worst hospital. For models 1 and 2, at least one hospital's dummy coefficient is significant in 22 and 21 cities, respectively. Including a third nest increases this figure to 24 cities. Figure 7 plots the estimated hospital dummy coefficients from models 2 and 4 for the two largest cities using the same order as in Figure 6. What stands out from the graph is the dispersion of estimated coefficients within each of the two cities to the point that both Barcelona and Madrid have hospitals ranked both among the top and the worst in Spain. These results strongly suggest that if coefficient estimates embody city-specific unobservable effects, they are unimportant.

7 Interpretation of Results

To evaluate our claim that physicians' choices are related to hospital quality, we perform two types of checks in this Section. First, we compare the rankings obtained from Models 1 to 4 with two alternative indicators of hospital quality: the risk-adjusted prevalence of nosocomial infections and the mortality rates for two different samples of about 40 hospitals each. Second, we regress the estimated hospital dummy coefficients on time-invariant geographical and hospital characteristics which may influence physicians' valuations.

7.1 Nosocomial infection data

Nosocomial infections are defined as infections which are neither present nor in incubation at the moment of admission into a hospital. Because these infections may depend on human factors, Gaynes (1997) claims that "surveillance of nosocomial infections can be used to assess the quality of care in the hospital." Since then, other researchers have used the incidence of nosocomial infections, among other measures, as proxy measures of quality of hospital care (for example, Weinstein, 1998, Wenzel, 1998, Navarrete-Navarro and Rangel-Frausto, 1999, Jarvis W.R., 2003). More recent research suggests that urinary tract nosocomial infections are the type of nosocomial infections that are most adequate to measure hospital quality (e.g. Stringham and Young, 2005, and Tinoco et al., 1994).

Since 1990, the Spanish Association for Preventive Medicine, Public Health and Higiene (Sociedad Española de Medicina Preventiva, Salud Pública e Higiene) assembles patient-level data on nosocomial infections collected by roughly 240 Spanish hospitals. To this date, this dataset constitutes the only source of homogeneous and comparable data on nosocomial infections in Spain. Regularly, the team that assembles the data publishes a report on the evolution of nosocomial infections in Spain (see Vaqué et al., 2003). These reports, however, do not disclose information either at the individual or at the hospital level. In order to access the individual level data for each hospital, we had to ask for permission to each hospital.

Not all hospitals with MIR vacancies participated in the EPINE study, and from those that did, only around 40 percent allowed us access to their confidential nosocomial infections data.[32] Our nosocomial infections dataset, therefore, matches only a subset of the hospitals in the physicians data.[33] As a result, only 39 hospitals from the physicians dataset can be matched with the nosocomial infections data.

From the nosocomial infections data, we have a total of $71,299$ patient observations from year 1995 to 2000. The average patient is 54 years old and 49% are female. The average number of nosocomial infections per patient is 0.084 and the average number of urinary track nosocomial infections per patient is 0.019.

When comparing the incidence of nosocomial infections across hospitals, there might be biases due to differences in the risk composition of patients across hospitals. On average, 62% of the patients have only a slight risk of contracting a nosocomial infection. Indeed 95.3% of these patients do not contract

[32] The low response rate and the difficulty in accessing this data, which pertains to the past, highlights the difficulties with using traditional methodologies in hospital care evaluation in Spain.

[33] Moreover, we disregard the observations from two hospitals due to problems with the codification of a variable concerning base risk of nosocomial infection.

any type of nosocomial infection while this figure is reduced to 88.3% for those at higher risk. Hence, in our exercise below, we restrict our sample to patients with moderate to severe risk of contracting a nosocomial infection. In spite of this restriction, hospitals are still heterogenous in terms of services, and patient characteristics. Table 9 in the Appendix shows basic descriptive statistics for the restricted sample. For example, the percentage of patients at each hospital with imunodeficiency ranges from zero to 24%, and the percentage of patients having any kind of surgery varies between 3.8% and 66%. Moreover, hospital size differs widely, from 97.9 patients at the smallest hospital to 1323 at the largest hospital.[34] Some studies highlight that hospital size is a good predictor of hospitals' rate of nosocomial infections. For example, Sax et al. (2002) show that even after adjusting for patient characteristics, hospital size is an extra risk factor for nosocomial infection (see also Mundy, 2002 for comments on this study).

Our variable of interest is the diagnosis of a urinary track nosocomial infection for two reasons. First, because as pointed out above, the literature has found evidence that this is the type of nosocomial infection more adequate to measure hospital quality; Second, we think that they are more dependent on human factors, related to more frequent procedures, and, hence, less correlated with patients' severity.[35] Because of the latter, endogeneity concerns that arise when using hospital outcome variables to measure quality, are potentially smaller.

The sample restricted to patients with moderate to severe risk of contracting a nosocomial infection has 25,559 observations. The average patient is now older, around 67 years of age, and 57% are male. Naturally, the average number of urinary nosocomial infections and nosocomial infections per patient are higher in this restricted sample, 0.03, and 0.14, respectively and they vary substantially across hospitals. The percentage of patients with some kind of chronic disease is also higher. For example, the percentage of patients in a state of coma is 5% as opposed to 2.4% in the overall nosocomial sample, and the percentage of patients with neoplasia is 35% as opposed to 16.3% (see Table 9 for statistics for other chronic diseases).

We use the nosocomial infections data to run several regressions at both patient and center-department-year level.[36] In all regressions, we obtain estimates of hospital dummy coefficients, which we use to construct alternative rankings of hospitals.

For the patient-level data, we run probit regressions on the diagnosis of a urinary track nosocomial infection for four different specifications: 1) unconditional, i.e. including only hospital dummies; 2) including hospital dummies and conditioniong on hospital size and size squared, where size is measured by the total number of patients per hospital-year; 3) in addition to 2), conditioning also on exogenous

[34]Hospital size is measured as the sum of all patients, not only those with moderate and high risk of contracting a nosocomial infection.

[35]For example, compared with all nosocomial infections, urinary nosocomial infections are less explained by the size of the hospital, the type of services in the hospital, as well as most of chronic diseases, and more related to patients' exogenous characteristics such as gender.

[36]Patient observations on the EPINE data are classified by center, department, and year. Most hospitals include a single center. There are, however, 4 large hospitals that include 3 to 4 centers.

regressors such as year dummies, number of patients per center-department-year and its squared, a female dummy, department dummies, dummies for main diagnosis, age and age squared; 4) in addition to 3) we control for variables such as dummies for all chronic diseases (see Table 9), a dummy for whether the patient had surgery, the duration until the patient contracts the first nosocomial infection, the number of non-nosocomial infections, the risk from the type of surgery in case the patient had surgery, dummies for the number of diagnosis (from 1 to 10), the existence of bed sores, and finally the gestational age, its squared and its cube.

For the center-department-year data, we run tobit specifications where the dependent variable is the number of urinary track nosocomial infections per patient at center-department-year level. Similarly to the probit regressions, we run tobits with the same four different variable specifications, where the regressors are averages at the center-department-year level.[37]

Table 6 shows Kendall's τ rank correlation statistic and simple correlations between the hospital rankings obtained using the physicians' data and the hospital rankings obtained using the urinary nosocomial infections data. For brevity, we comment only the correlations for two specifications using the physicians data (rows) versus four specifications using the urinary nosocomial infections data (columns). Columns refer to the different probit and tobit regressions. In the first row, we show the correlations with "Sample" i.e. the average across years of the order of the first physician to choose a given hospital (any specialty). This model may be interpreted as an unconditional specification for the physicians' data. The second row shows the correlations with the hospital dummy coefficients (multiplied by -1) obtained from our Model 4.

[37] Results for the patient and center-department-year regressions are available from the authors upon request.

	Comparison of Rankings based on Physicians' data with Rankings based on urinary nosocomial infections (EPINE data) **Kendall's τ rank correlation**							
	Probit				**Tobit (b)**			
	uncond.	size	exog.	all	uncond.	size	exog.	all
Sample	0.144 (0.100)	0.468 (0.000)	0.363 (0.001)	0.350 (0.001)	−0.069 (0.727)	0.485 (0.000)	0.471 (0.000)	0.487 (0.000)
Model 4	0.180 (0.055)	0.455 (0.000)	0.382 (0.000)	0.347 (0.001)	0.074 (0.257)	0.574 (0.000)	0.479 (0.000)	0.431 (0.000)
	Simple correlation							
	Probit				**Tobit (b)**			
	uncond.	size	exog.	all	uncond.	size	exog.	all
Sample	0.137 (0.202)	0.606 (0.000)	0.464 (0.001)	0.427 (0.003)	−0.049 (0.617)	0.593 (0.000)	0.604 (0.000)	0.619 (0.000)
Model 4	0.220 (0.089)	0.641 (0.000)	0.501 (0.001)	0.503 (0.001)	0.114 (0.245)	0.691 (0.000)	0.670 (0.000)	0.631 (0.000)

Notes: (a) p-values for the test of independence between the rankings against the null of a positive correlation are in parenthesis. These are not corrected for the fact that rankings are estimates.
(b) correlations for the tobit models are based on tobit regressions per year-center-department
(c) hospital dummy estimates for Model 4 were premultiplied by (-1)

Table 6: Hospital dummy coefficients: physicians vs. nosocomial data

Table 6 deserves several comments. The correlations between the hospital dummy coefficients from the physicians data and the coefficients from the nosocomial data are positive and strongly statistically significant for all conditional specifications. Based on simple correlations, Model 4 provides rankings which are generally more positively correlated with rankings using nosocomial infections data. Based on Kendall's rank correlations, both Sample and Model 4 ranks are similarly correlated to the estimates from the conditional probits and tobits. Finally, controlling for variables in addition to *size*, generally reduces correlations because these controls capture hospital characteristics that physicians perceive as quality, such as department fixed effects and dummies for type of diagnoses. As an illustration, figure **??** plots the hospital dummy coefficients from Model 4 versus the hospital dummy coefficients from all specifications for the tobit models.

These conclusions should, however, be interpreted with caution for several reasons. First, our results using both datasets rely on a limited number of controls. Although, we think nosocomial urinary infections suffer less from a selection bias than other output measures and are, therefore, more appropriate to measure quality of care, there may still be some unobserved severity correlated with the hospital dummies that induces some bias in the rankings from the nosocomial infection data. Second, our nosocomial infections data is restricted to only 41 hospitals, a small sample which may reduce the precision of the correlation statistics. Moreover, not all hospitals with MIR vacancies participate in the

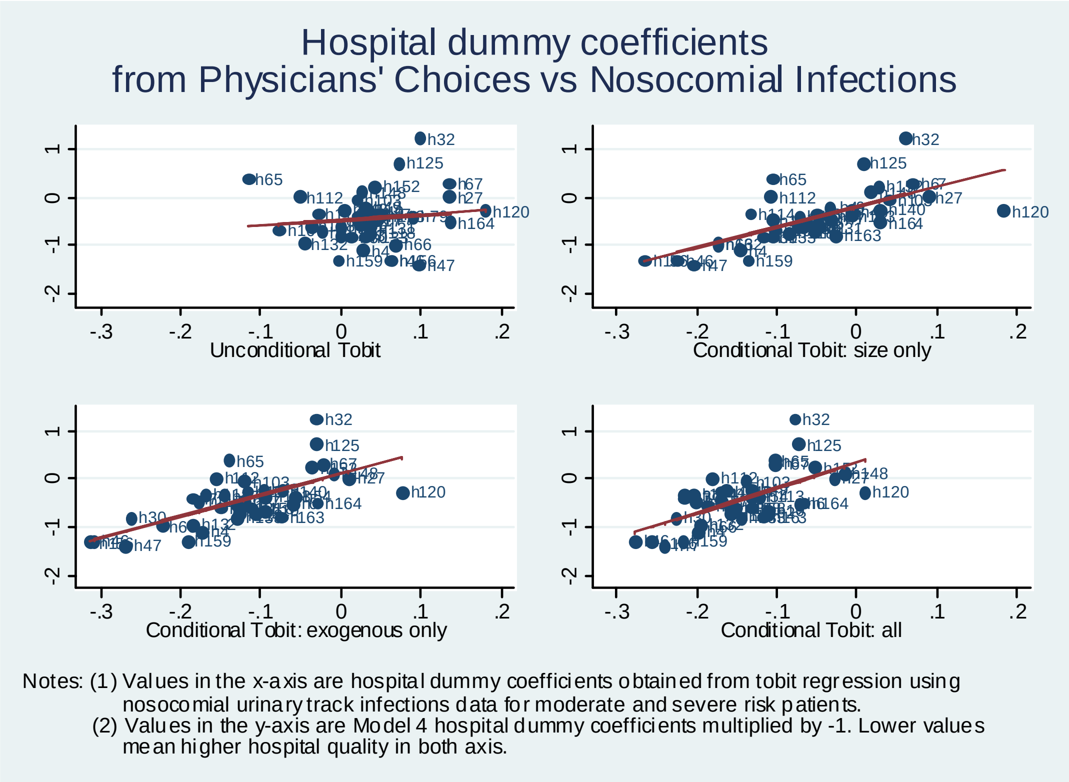

Hospital dummy coefficients
from Physicians' Choices vs Nosocomial Infections
Unconditional Tobit
Conditional Tobit: size only
Conditional Tobit: exogenous only
Conditional Tobit: all
Notes: (1) Values in the x-axis are hospital dummy coefficients obtained from tobit regression using nosocomial urinary track infections data for moderate and severe risk patients.
(2) Values in the y-axis are Model 4 hospital dummy coefficients multiplied by -1. Lower values mean higher hospital quality in both axis.

Figure 8:

EPINE study, and from those which do, only about 40% gave us access to their confidential nosocomial infections data. Figure 9 compares distributions of hospital dummy estimates for models 2 and 4 according to the hospital's participation in the EPINE study and whether we were granted permission to use its nosocomial data. Panels at the left side of the graph show that hospitals participating in the EPINE study have higher hospital dummy estimates. Indeed, if we look at all hospitals that enter the EPINE study and have MIR vacancies, the average of the hospital dummy estimates from models 2 and 4 are 0.49 and 0.53, respectively, while for those hospitals not participating in the EPINE study, the averages are 0.40 and 0.37, respectively. We interpret this result as highlighting that hospitals in the EPINE study are, on average, of better quality. Therefore, results shown in Table 6 are still valid but they only refer to relatively higher-quality hospitals. Next, we should compare hospital dummy estimates for those hospitals that gave us permission to use their nosocomial data with the estimates for those that did not. For the former group, the averages are surprisingly lower, 0.40 and 0.45, than for the latter, 0.55 and 0.58, partially offsetting the effect on average quality from participation in EPINE. Moreover, the support of the distributions of hospital dummy estimates is similar for the two samples, as shown by the right hand side graphs in figure 9.

7.2 Mortality data

As discussed in Section 2, hospitals' disease-specific mortality rates have been used to assess hospital quality. In Spain there is no publicly available data on hospital specific mortality.[38] However, some hospitals publish in their annual reports the aggregate rate for the hospital and/or the departamental-specific rates.

After contacting all hospitals from the physicians dataset, we were able to gather at least one annual report from 79 hospitals, which represents less than 45% of the hospitals. From the reports available, only 40 hospitals published information on aggregate mortality (defined as deaths over admissions) for at least one of the six years of the sample. We drop the psiquiatric hospitals from the mortality data, hence our mortality data contains information on 39 hospitals and 166 observations.

We run linear regressions of hospital-year mortality odds ratios for three different specifications: 1) unconditional, i.e. including only hospital dummies; 2) including hospital dummies and conditioniong on annual admissions; 3) in addition to 2), conditioning also on average length of stay (LOS).

Table 7 shows Kendall's τ rank correlation statistic and simple correlations between the rankings obtained using the physicians' data and those using the mortality data. As in Table 6, in the first row, we show the correlations with "Sample", and the second row shows the correlations with the hospital dummy coefficients (multiplied by -1) obtained from our Model 4.

[38] There is a rich but politically sensitive dataset called CMBD, after the Spanish version for "Data Base Minimum Set", which contains this information. Access to this dataset is usually not granted for research purposes, and we were unsuccessful in our efforts to obtain it.

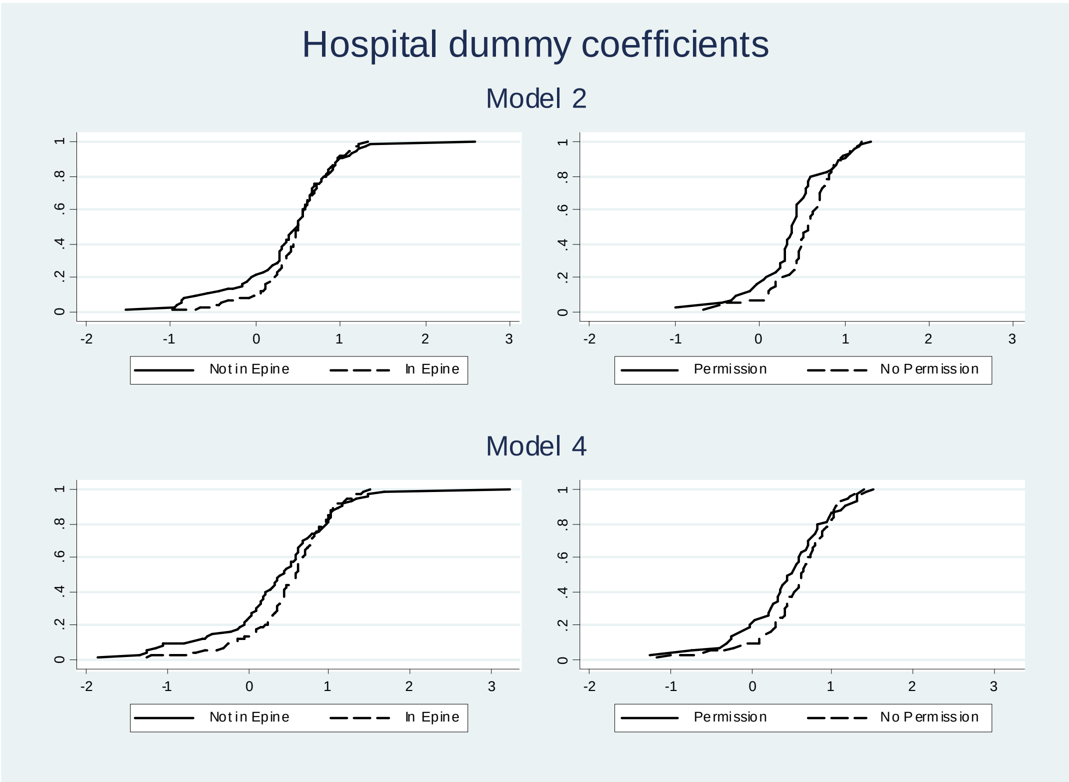

Hospital dummy coefficients
Model 2
Not in Epine
In Epine
Permission
No Permission
Model 4
Not in Epine
In Epine
Permission
No Permission

Figure 9:

	Rankings based on hospital level mortality data			
	Kendall's τ rank correlation			
	uncond.	admissions only	LOS only	admissions and LOS
Sample	-0.090 (0.788)	-0.439 (1.000)	0.054 (0.318)	0.159 (0.078)
Model 4	0.028 (0.404)	-0.304 (0.997)	0.162 (0.075)	0.273 (0.008)
	Simple correlation			
	uncond.	admissions only	LOS only	admissions and LOS
Sample	0.044 (0.395)	-0.489 (0.992)	0.191 (0.123)	0.283 (0.041)
Model 4	0.089 (0.295)	-0.429 (0.997)	0.257 (0.057)	0.372 (0.010)

Notes: (a) p-values for the test of independence between the rankings against the null of a positive correlation are in parenthesis.
(b) correlations not corrected for the fact that rankings are estimates.
(c) hospital dummy estimates for Model 5 were premultiplied by (-1)

Table 7: Hospital dummy coefficients: physicians vs. mortality data

The main conclusion from Table 7 is that only after controlling for LOS is the correlation between the ranking from Model 4 and rankings using mortality data significantly positive. As an illustration, figure 10 plots the hospital dummy coefficients from Model 4 versus the hospital dummy coefficients from all specifications for the regression models.

The correlation between the risk-adjusted mortality rates and our ranking could be underestimated for several reasons. First, the published mortality dataset only allows us to control for admissions and average length of stay. Other commonly used time-variant risk adjusters, such as case-mix, are ommited, which may lead to biased estimates of the hospital fixed effects in the mortality equation. Second, hospital fixed effects in the mortality equation may be picking up time-invariant health characteristics in the population served by the hospital that are not related to the hospital's quality. Finally, there is a potential selection problem in the mortality dataset. If only the best hospitals publish average mortality rates, the sample variation in mortality rates may not be sufficient to pick up a significant correlation with our hospital dummy estimates.

As in the previous section we check for potential selection bias using the estimated hospital dummy coefficients from models 2 and 4. Figure 11 shows the distribution of estimates of hospital dummy coefficients by availability of annual reports and mortality rates. The figure suggests that there is truncation and selection both in the availability of reports and in the decision to publish the mortality rates. First, if we look at the left hand-side graphs, the range of values for those hospitals for which

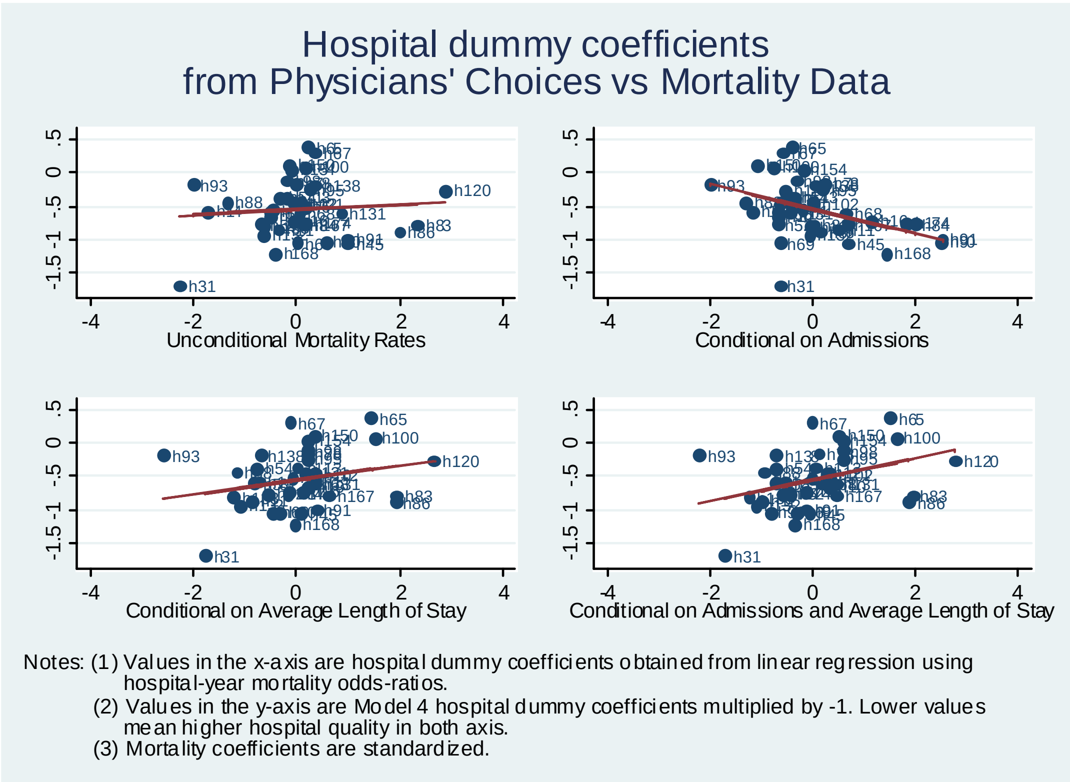

Hospital dummy coefficients
from Physicians' Choices vs Mortality Data
Unconditional Mortality Rates
Conditional on Admissions
Conditional on Average Length of Stay
Conditional on Admissions and Average Length of Stay
Notes: (1) Values in the x-axis are hospital dummy coefficients obtained from linear regression using hospital-year mortality odds-ratios.
(2) Values in the y-axis are Model 4 hospital dummy coefficients multiplied by -1. Lower values mean higher hospital quality in both axis.
(3) Mortality coefficients are standardized.

Figure 10:

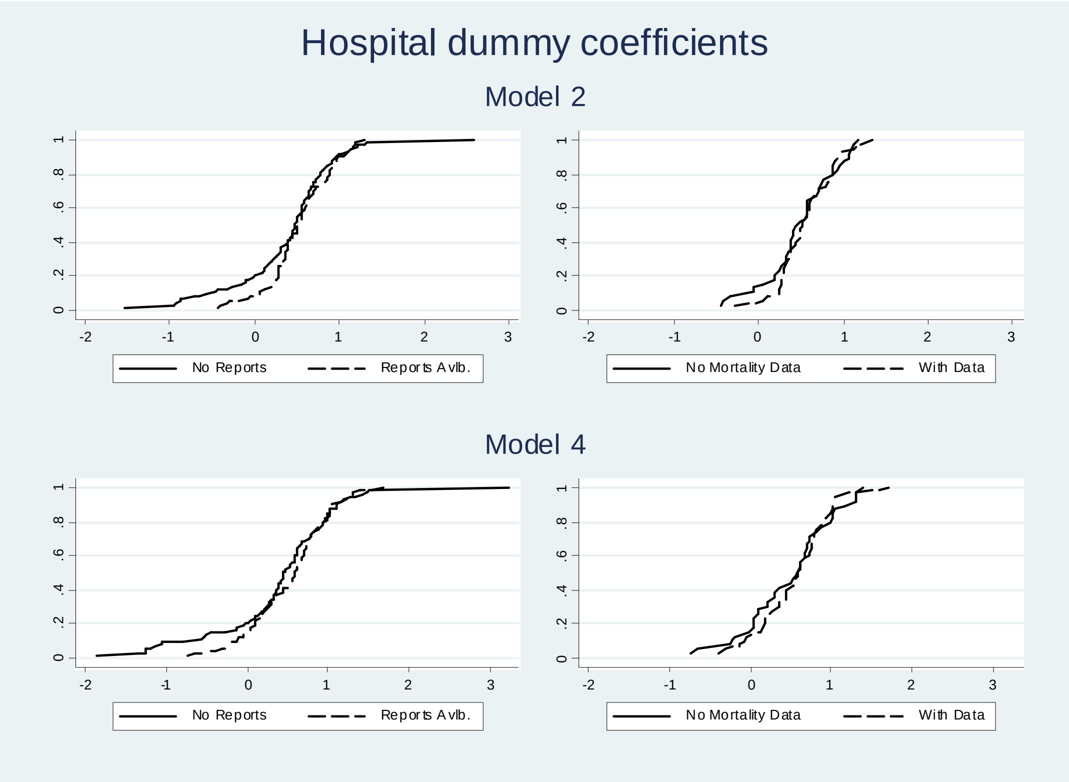


Figure 11:

at least one report is available is smaller than the range for those for which no report is available. Moreover, for the latter group the average value of the estimated coefficients for models 2 and 4 is lower (0.39 and 0.40) than the average for the former (0.56 and 0.57).

Second, many factors may influence the decision to publish mortality data. Presumably, those hospitals with the highest mortality rates will be less likely to disclose this information to the general public. If we compare the two right hand-side graphs, it is clear that those with the lowest quality, among those with reports, decide not to publish the mortality rates. It is also true, although not clear from the graphs, that those with the highest coefficient estimates do publish their mortality data. Nonetheless, the two distributions are similar in the common support.

Despite reasons to believe that the correlation between the risk-adjusted mortality rates and our ranking is underestimated, we still find a strong positive correlation.

To sum up, we observe a positive correlation between the risk-adjusted mortality rates and our ranking, and this observed correlation is likely underestimated.

7.3 Correlations with Location Characteristics

In this subsection we check wether hospital dummy coefficient estimates correlate with time-invariant location characteristics unrelated to hospital quality. Table **??** shows regression results using as dependent variables the estimated hospital dummy coefficients from Models 2 and 4. We include three sets of controls which attemp to reflect quasi-permanent differences in quality of life across the country. The first set refers to variables for which we have major regional-level information (Autonomous Community).[39] It includes the average number of people per dwelling (ppdwell), the percentage of dwellings which are owned by the occupants (own), and percentage of dwellings with self-reported problems with pollution and other environmental problems (polute), and the percentage of people with higher education (higheduc). The second set of data refers to controls for which provincial-level information is available. It includes the number of court decisions per thousand people (court), the number of divorces per thousand people (divorce), the number road traffic accidents per thousand people (accident), the number of suicides in the provincial capital per $100,000$ people (suicide), the number of days with average temperature below 0^oC (below0), and above 25^oC (above25), and the percentage of tourist-days per resident-days in a year (tourism).[40]. Finally, we also use published results from a survey carried out in 2006 by the Spanish consumers' association OCU on different aspects of city quality of life (OCU, 2007). The survey covers 17 cities and their metropolitan areas, where 59% of all hospitals are located. Controls from this survey include average respondents valuations from one to 10 about the city's culture, labour market, crime, and availability of health, educational, and other city services.

[39] The source of these data is the webpage of the Spanish Statistical Office (www.ine.es). Due to lack of availability, we use data from 2004, a reasonable proxy for the expectations of the physicians in our sample period.

[40] Road traffic data was collected from the Road Accidents Annual Report for the years 1998 to 2000 published by the Spanish Department of Transportation. The source for the other controls is the Spanish Statistical Office for the year 2000 except for the temperature variables, for which we take the average for the 1997:2000 period.

	Model 2				Model 4			
	coef	s.d.	coef	s.d.	coef	s.d	coef	s.d.
ppdwell	0.527	0.538	−0.783	2.605	0.710	0.668	−0.252	3.225
own	0.014	0.022	−0.006	0.0738	0.048	0.027	0.064	0.091
polute	0.020	0.018	0.052	0.098	0.042	0.022	0.019	0.121
higheduc	−0.005	0.014	0.073	0.081	−0.020	0.018	0.053	0.100
court	0.039	0.087	−0.370	0.475	0.033	0.108	−0.375	0.588
divorce	−0.144	0.363	*n.i.*	*n.i*	0.261	0.450	*n.i.*	*n.i*
tourism	0.024	0.025	−0.103	0.106	0.019	0.031	−0.046	0.131
accident	−0.122	0.086	0.446	0.424	−0.070	0.106	0.538	0.525
below0	−0.000	0.003	0.041	0.046	−0.003	0.003	0.014	0.057
above25	−0.002	0.001	0.008	0.007	−0.002	0.002	0.010	0.009
suicide	0.002	0.003	0.004	0.031	−0.001	0.003	0.007	0.038
culture	–	–	−1.434	1.600	–	–	−1.183	1.981
lab. makt	–	–	−0.967	0.795	–	–	−0.861	0.984
services	–	–	2.957	3.050	–	–	2.184	3.776
crime	–	–	1.069	1.085	–	–	0.825	1.343
education	–	–	−2.440	1.880	–	–	−1.281	2.328
health	–	–	1.226	1.221	–	–	0.648	1.511
constant	−1.95	3.22	−8.56	13.58	−5.61	4.00	−13.44	16.81
obs	183		108		183		108	
R^2	0.070		0.090		0.090		0.092	
Note: n.i. means not identified								

Table 8: Kendall's tau rank correlation between hospital rankings

Results show that at most we can explain 9.2% of the variation in hospital dummy estimates. No control is significant at the 5% level. In some specifications, a couple of controls have p-values around 0.07, but these values increase when all other non-significant variables are dropped from the regressions. To sum up, time-invariant location characteristics cannot explain differences in the estimated hospital dummy coefficients.

8 Conclusions

In this paper, we propose an alternative methodology based on a revealed preferences argument to rank hospitals. More specifically, we use Spanish data on the choices over hospital vacancies made by young physicians at the beginning of their specialized training. In Spain, after graduation from medical school, physicians choose hospital training vacancies sequentially depending on merit.

We model the physicians' decisions as a nested logit that takes into account, among other controls, preferences for geographical proximity. We construct hospital rankings using the estimates of hospital dummy coefficients. We find significant differences in physicians' valuations of hospitals. These differences are robust to the introduction of hospital and regional controls as well as the introduction of a third nest that accounts for potential language barriers within the national system. We find that physicians value geographical and language proximity. Moreover, there are significant differences in the way different specialties are valued by male and female physicians.

We show that physicians' choices are related to hospital quality. First, observable time-invariant city characteristics are unrelated to our estimates of hospital dummy coefficients. Second, we show that our ranking is correlated to rankings obtained using data on risk-adjusted prevalence of nosocomial infections and mortality rates for two different subsamples of hospitals.

Although we apply our methodology using Spanish data, our results convey the relevance of exploiting physicians' labour market data, possibly in conjunction with other data, to assess relative hospital quality in all countries.

References

[1] Ackerberg, Daniel A., Matilde P. Machado and Michael H. Riordan (2006): "Benchmarking for Productivity Improvement: a Health Care Application," International Economic Review, vol 47, n^{o}1, pp 161-201.

[2] Allsopp, A. and J. D. Hey, (2000): "Two Experiments to Test a Model of Herd Behaviour", Experimental Economics, 3(2):121-136.

[3] Avery, Christopher, Mark Glickman, Caroline Hoxby, Andrew Metrick (2005): "A Revealed Preference Ranking of U.S. Colleges and Universities," mimeo.

[4] Banerjee, A. V. (1992): "A simple model or herd behaviour", Quartely Journal of Economics, 107, 797-817.

[5] Bradlow, Eric T. and Peter S. Fader (2001): "A Bayesian Lifetime Model for the Hot 100 Billboard Songs," Journal of the American Statistical Association, Vol. 96, No. 454, Applications and Case Studies, pp. 368:81.

[6] Burgess, Simon, Gossage, Denise, and Propper, Carol (2003): "Explaining Differences in Hospital Performance: Does the Answer Lie in the Labor Market," CEPR DP 4118.

[7] Chapman, Randall G. and Staelin, Richard (1982): "Exploitingh Rank Ordered Choice Set Data Within the Stochastic Utility Model," Journal of Marketing Research, vol. XIX (August 1982), pp. 288-301.

[8] Commons, Margareth, McGuire, Thomas G and Michael Riordan: "Performance Contracting for Substance Abuse Treatment" Health Services Research, 32(5) December 1997 pp 630-650

[9] Dranove, David, Kessler, Daniel, McLellan, Mark, and Satterthwaite, Mark (2003): "Is more Information Better? The Effects of "Report Cards" on Health Care Providers," The Journal of Political Economy, 111 (3), 555-588.

[10] Gaynes RP. (1997) Surveillance of nosocomial infections: a fundamental ingredient for quality, Infect Control Hosp Epidemiology, Jul; 18(7):475-8.

[11] González, Beatriz, Patricia Barber, M^a Isabel Tocino, Elena Rodriguez (1997): "El mercado laboral sanitario y sus consecuencias en la formación. *Numerus Clausulus*". Mimeo Universidad de las Palmas de Gran Canaria.

[12] González, Beatriz (2000): "Formación y empleo de profesionales sanitarios en España. Un análisis de desequilibrios," Gaceta Sanitaria, 14(3), pp 237-246.

[13] González, Beatriz, Patricia Barber (2004): "El programa MIR como innovación y como mecanismo de asignación de recursos humanos." Mimeo

[14] Gowrisankaran, Gautam and Town, Robert (1999): "Estimating the Quality of Care in Hospitals Using Instrumental Variables" Journal of Health Economics, 18: 747-67.

[15] Geweke, John, Gowrisankaran, Gautam and Town, Robert (2003): "Bayesian Inference for Hospital Quality in a Selection Model," Econometrica 71, pp 1215-1238.

[16] Heckman, J. (1979): "Sample selection as a specification error," Econometrica 47, 153-161.

[17] Herriges, J. and C. Kling (1996): "Testing the consistency of nested logit models with utility maximization," Economic Letters 50, 33–39.

[18] Holl, Adelheid (2004): "Manufacturing location and impacts of road transport infrastructure: empirical evidence from Spain", Regional Science and Urban Economics 34 (3), 341-363.

[19] Kling, C. and J. Herriges (1995): "An empirical investigation of the consistency of nested logit models with utility maximization", American Journal of Agricultural Economics 77, 875–884.

[20] Lu, Mingshan (1999): "Separating the "True Effect" from "Gaming" in Incentive Based Contracts in Health Care," Journal of Economics and Management Strategy, 8, issue 3, 383-431.

[21] Maxwell RJ. (1984): "Quality assessment in health," British Medical Journal; 288: 1470-1472.

[22] McLellan, Mark and Staiger, Douglas (1999a): "Comparing Hospital Quality at for-profit and not-for-profit hospitals," NBER WP 7324.

[23] McLellan, Mark and Staiger, Douglas (1999b): "The Quality of Health Care Providers" NBER WP 7327.

[24] McFadden, Daniel (1974): "Conditional Logit Analysis of Qualitative choice Behavior," in Frontiers of Econometrics, P. Zarembka, ed. , New York, Academic Press, pp105-142.

[25] McFadden, Daniel (1978): "Modelling the choice of residential location" in A. Karlquist, et al. (eds.), Spatial Interaction theory and planning models, Amsterdam, North-Holland.

[26] Moscucci, Mauro, Eagle, Kim A., Share, David, Smith, Dean, De Franco, Anthony C. , O'Donnel, Michael, Kline-Rogers, Eva, Jani, Sandeep M., Brown, David L. (2005): "Public Reporting and Case Selection for Percutaneous Coronary Interventions: An analysis from two large multicenter Percutaneous Coronary Interventions databases," Journal of the American College of Cardiology, vol. 45 No. 11, pp 1759-65.

[27] Mundy, Linda M. (2002): “Does Hospital Size Predict Risk for Nosocomial Infection?”, Journal Watch Infectious Diseases.

[28] Navarrete-Navarro S. and Rangel-Frausto MS (1999): “Las infecciones nosocomiales y la calidad de la atencion médica”. Salud Publica Mex, 41 suppl 1:S64-S68.

[29] Normand, Sharon-Lise T., Glickman, Mark E., and Gatsonis, Constantine A. (1997) “Statistical Methods for Profiling Providers of Medical Care: Issues and Applications”, Journal of American Statistical Association, Vol. 92, No. 439, Applications and Case Studies, 803-814.

[30] OCU (2007): “Encuesta sobre la calidad de vida de las ciudades”, OCU-COMPRA MAESTRA, Vol 317, 29-33.

[31] Roth, Alvin E. and Sotomayor, Marilda (1990): “Two-Sided Matchings Markets: A Study in Game-Theoretic Modeling and Analysis”. Econometric Society Monograph. Cambridge, MA: Cambridge University Press.

[32] Roth, Alvin E. (1984), “The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory,” Journal of Political Economy, 92, pp991-1016.

[33] Roth, Alvin E. (1989), “A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians and Surgeons in the U.K.,” American Economic Review, vol. 81, June 1991, pp415-440

[34] Sax, H., Pittet, D. and the Swiss-NOSO Network (2002): “Interhospital Differences in Nosocomial Infection Rates: Importance of Case-Mix Adjustment”, Arch Intern Med.;162:2437-2442.

[35] Shahian, David M., Normand, Sharon-Lise, Torchiana, David F., Lewis, Stanley M., Pastore, John O., Kuntz, Richard E. and Dreyer, Paul I. (2001): “Cardiac Surgery Report Cards: Comprehensive Review and Statistical Critique,” The Annals of Thoracic Surgery, Vol. 72, pp 2155-2168.

[36] Shen, Yujing (1999), “Selection and its impacts on Health Payment Systems,” PhD dissertation, Boston University.

[37] Stringham J., Young, N. (2005): “Using MedPAR Data as a Measure of Urinary Tract Infection Rates: Implications for the Medicare Inpatient DRG Payment System”, Perspectives in Health Information Management, 2:12.

[38] Hofstede, Frenkel and Michel Wedel (1998): “A Monte Carlo Study of time-aggregation in continuous-time and discrete time parametric hazard models,” Economics Letters, 58, pp149-156.

[39] Jarvis WR. Benchmarking for prevention: the Centers for Disease Control and Prevention’s National Nosocomial Infections Surveillance (NNIS) system experience, Infection. 2003 Dec;31 Suppl 2:44-8.

[40] Tinoco J.C., M.C., Zequiel Hernandez-Ruiz, M.C., Jaime Salvador-Moysen, M.I.S.P., Irma Rivera-Morales, M.C.(1994) “Infecciones Nosocomiales de vias urinarias en un Hospital de segundo nivel”, Salud Publica De Mexico, 36:1.

[41] Train, Kenneth, (2003): Discrete Choice Methods with Simulation, Cambridge University Press.

[42] Train, K., D. McFadden, and M. Ben-Akiva (1987), “The demand for local telephone service: A fully discrete model of residential calling patterns and service choice”, Rand Journal of Economics 18, 109123.

[43] Vaqué (2003) "Evolución De La Prevalencia De Las Infecciones Nosocomiales En Los Hospitales Españoles, Epine 1990-2003" J. Vaqué (ed). Sociedad Española de Medicina Preventiva, Salud Pública e Higiene.

[44] Wen, C. H. and F.Koppelman (2001): "The generalized nested logit model", Transportation Research B 35, pp 627–641.

[45] Wenzel RP. Hospital epidemiology. Beyond infection control and toward quality assurance. Clin Microbiol Newsletter 1998;10:60-62.

[46] Weinstein, R. A. Nosocomial Infection Update, Emerging Infectious Diseases, Vol. 4, No. 3, July–September 1998

9 Data Appendix

In this appendix we give some of the statistical evidence that guided some of the specification choices made. In particular the next table shows that most people choose a vacancy in the same region as their college. Each column represents a college location, so for example 79.05 percent of the students that graduated in Andalucian colleges choose a hospital located in Andalucia. The numbers in diagonal represent the percentage of physicians who choose a hospital in the same region as their college.

	Patient level data					Hospital averages				
	obs	mean	s.d.	min	max	obs	mean	s.d.	min	max
No. of urinary nosoc. infections	25559	0.03	0.174	0	3	41	0.033	0.013	0.009	0.069
No. of nosoc. infections	25559	0.140	0.419	0	4	41	0.139	0.039	0.045	0.217
Female	25358	0.43	0.50	0	1	41	0.436	0.039	0.325	0.539
Age	25436	66.55	18.72	0.003	99	41	68.44	5.11	53.57	79.48
Patients per hospital-year (*size*)	—	—	—	—	—	41	382.9	278.2	97.93	1322.7
Patients per center-department-year	—	—	—	—	—	41	54.3	27.6	19.8	159.9
No. of non-nosoc. infections	25559	0.242	0.552	0	4	41	0.263	0.101	0.06	0.569
Days until 1st nosoc. infection	24896	16.25	44.93	1	1498	41	15.76	4.71	9.18	32.67
Diagnosis per patient	24842	2.32	1.40	1	9	41	2.368	0.532	1.19	3.45
Surgery	25239	0.281	0.449	0	1	41	0.285	0.115	0.038	0.660
State of coma	24561	0.052	0.222	0	1	41	0.051	0.026	0.005	0.104
Kidney failure	24488	0.124	0.330	0	1	41	0.118	0.041	0.018	0.228
Diabetes	24586	0.221	0.415	0	1	41	0.229	0.047	0.133	0.326
Neoplasia	24493	0.352	0.477	0	1	41	0.353	0.114	0.062	0.629
Chronic lung disease	24542	0.186	0.389	0	1	41	0.199	0.080	0.076	0.465
Imunodeficiency	24532	0.093	0.291	0	1	41	0.093	0.050	0	0.242
Neutropenia	24500	0.039	0.194	0	1	41	0.035	0.020	0	0.082
Cirrhosis	24512	0.044	0.205	0	1	41	0.044	0.019	0.002	0.101
Drug addiction	24508	0.018	0.133	0	1	41	0.021	0.018	0	0.084
Obesity	24477	0.112	0.315	0	1	41	0.111	0.051	0.018	0.218
Malnutrition	25530	0.095	0.293	0	1	41	0.092	0.052	0	0.246
Bed sores	24170	0.065	0.247	0	1	41	0.071	0.041	0.014	0.256
Fetal age	25559	0.237	2.79	0	42	41	0.151	0.228	0	0.809
Risk from surgery	25239	0.412	1.02	0	5	41	0.496	0.276	0.021	1.340

Table 9: Basic Statistics for the EPINE subsample of moderate to severe risk patients