Patient Mobility, Health Care Quality and Welfare

Kurt R. Brekke* Rosella Levaggi† Luigi Siciliani‡ Odd Rune Straume§

August 15, 2011

Abstract

Patient mobility is a key issue in the EU who recently passed a new law on patients' right to EU-wide provider choice. In this paper we use a Hotelling model with two regions that differ in technology to study the impact of patient mobility on health care quality, health care financing and welfare. A decentralised solution without patient mobility leads to too low (high) quality and too few (many) patients being treated in the high-skill (low-skill) region. A centralised solution with patient mobility implements the first best, but the low-skill region would not be willing to transfer authority as its welfare is lower than without mobility. In a decentralised solution, the effects of patient mobility depend on the transfer payment. If the payment is below marginal cost, mobility leads to a 'race-to-the-bottom' in quality and lower welfare in both regions. If the payment is equal to marginal cost, quality and welfare remain unchanged in the high-skill region, but the low-skill region benefits. For a socially optimal payment, which is higher than marginal cost, quality levels in the two regions are closer to (but not at) the first best, but welfare is lower in the low-skill region. Thus, patient mobility can have adverse effects on quality provision and welfare unless an appropriate transfer payment scheme is implemented.

Keywords: Patient mobility; Health care quality; Regional and global welfare.

JEL Classification: H51; H73; I11; I18

*Corresponding Author: Department of Economics and Health Economics Bergen, Norwegian School of Economics, Helleveien 30, N-5045 Bergen, Norway. E-mail: kurt.brekke@nhh.no.

†Department of Economics, University of Brescia, Via San Faustino 74b, 25100 Brescia, Italy. E-mail: levaggi@eco.unibs.it.

‡Department of Economics and Centre for Health Economics, University of York, Heslington, York YO10 5DD, UK; and C.E.P.R., 90-98 Goswell Street, London EC1V 7DB, UK. E-mail: ls24@york.ac.uk.

§Department of Economics/NIPE, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal; and HEB, Department of Economics, University of Bergen. E-mail: o.r.straume@eeg.uminho.pt.

1 Introduction

Cross-border patient mobility is a key issue in the European Union at the moment. Despite the fact that patients in EU member states are allowed to seek health care in other EU countries, patient mobility is still very low, especially for planned health care treatments.[1] A natural explanation for low mobility is that patients prefer to be treated in their home country. However, there might be other causes. Patients might be denied access and/or reimbursement if they demand treatment in a foreign EU country.[2] In March 2011 the EU council passed a new law that gives citizens in EU countries the right to choose among health care providers across all EU member states.[3] The new law intends to limit the scope for EU countries (or providers within EU countries) to deny foreign EU citizens access to their health care provision. The law also explicitly states that EU countries cannot refuse to reimburse patients who seek cross-border medical treatment when this treatment is covered in their home country.[4] Thus, by lowering important barriers for patients seeking care in another EU country, the new law is likely to stimulate patient mobility across EU member states.

In this paper we ask whether patient mobility is desirable or not from a welfare perspective. Clearly, the answer to this question relies on what are the effects of patient mobility on the provision and financing of health care within each country, which is what we will study in detail. While our paper is motivated by the on-going debate and the new legislation in the EU on cross-border medical treatment, our analysis also applies to patient mobility within country borders, where regions are separate jurisdictions. For example, Sweden has a decentralised health care system, which is financed primarily through taxes levied by county councils and municipalities. County councils also regulate the level of service offered by the providers. In 2003 a 'free choice reform' was implemented, which allows patients to apply for health care outside their home county, though needing to pay out-of-

[1] According to the EU Commission (2006) the demand for cross-border health care represents only around 1% of public spending on health care, which is currently around €10 billion. This estimate includes cross-border health care which patients had not planned in advance (such as emergency care), which means less than 1% of the expenditure and movement of patients is for planned cross-border health care, like hip and knee operations or cataract surgery.

[2] Several EU Court cases illustrate the problem where patients are refused reimbursement by the home country for cross-border treatment; see, e.g., Case C-158/96 [Kohll, 1998], Case C-120/95 [Decker, 1998] and Case C-372/04 [Watts, 2006]. Although the EU Court decided in favour of the patients, it is still likely that patients face uncertainty and costs related to reimbursement for cross-border treatment. See, e.g., the EU commision (2006) for a discussion of these cases.

[3] Directive 2011/24/EU of the European Parliament and of the Council of 9 March 2011 on the application of patients' rights in cross-border healthcare.

[4] The EU directive (chapter III) defines some basic principles for the cross-border reimbursement, but is not very specific on the transfer payments across the member states and the reimbursement to patients seeking cross-border care. Thus, the EU member states have some discretion in designing the reimbursement rules.

pocket for the extra travel costs. The home county would need to compensate the county providing the treatment to their residents. Similarly, in Italy each Region is responsible for the provision of health care. However, many patients seek care in a different Region from the one where they reside and a system of transfers is in place: 'importing' Regions are compensated on the basis of the number of patients treated from the 'exporting' ones. In Canada, Provinces are responsible for the provision of health care. Mobility across Provinces is generally limited to emergency and sudden illness or allowed only in special circumstances (for example a specialised treatment not offered in a Province) under prior approval.

Relatively little is known and understood about patient mobility and its consequences for health care provision, health care financing and regional and global (inter-regional) welfare. We aim to contribute towards filling this gap in the literature. In order to analyse patient mobility across separate jurisdictions, we make use of a Hotelling model with two regions. Health care is financed through income taxation. Patients receive care for free at the point of consumption, but face the cost of travelling to the provider for treatment. The policy makers in each region decide on the quality of health care provision in their region and the corresponding tax rate to finance their health care expenditures. The regions are identical except for their technology in providing health care quality, e.g., due to access to more skilled doctors, better medical technology, better facilities, etc. All else equal, the high-skill region will offer higher health care quality than the low-skill region. This is the source of patient mobility in our model.

The objective of our study is twofold. First, we compare the decentralised system with no mobility (for example the old system within the EU) with a centralised one. We show that a centralised solution implements the first best. Although we do not envisage the EU taking over the funding of health care systems directly, the centralised solution remains a useful benchmark as total welfare is highest under this solution: it coincides with the first best. Second, and most importantly, we compare the decentralised system with no mobility (again the old system within the EU) with a decentralised system where mobility is allowed and (potentially) a system of transfers can be put in place (the new system within the EU).

Centralisation versus decentralisation with no mobility. Compared to a decentralised system with no mobility, a centralised provision of quality which allows mobility is welfare improving. Since the two regions differ in their quality, the patient at the border between the high- and low-skill region

is willing to travel further to obtain the extra quality of care. Therefore a decentralised system with no mobility implies that the high-skill (low-skill) region treats too few (too many) patients. It also implies too low quality in the high-skill region and too high quality in the low-skill region. Since demand is higher (lower) for the high-skill (low-skill) region under the centralised solution, the optimal quality is higher (lower). In the high-skill region patients are better off under a centralised system since they receive higher quality. In the low-skill region the effect on patients' utility from health gains is mixed: patients who travel from the low-skill to the high-skill region benefit from the higher quality provided in this region, but patients who stay in the low-skill region have lower quality compared to a decentralised solution. Whether each region pays more or less taxes under a centralised solution is in general indeterminate. However, we show that if the cost of quality provision takes a quadratic form, health expenditures are higher in the high-skill region and it benefits from an implicit subsidy from the low-skill region so that the high-skill (low-skill) region is overall better (worse) off under centralisation.

The result that allowing mobility under a centralised solution is welfare improving does not necessarily imply that it is welfare improving under a decentralised one. It however shows that it has the potential to improve welfare. As we discuss below, whether mobility increases or reduces welfare depends on the system of transfers between the different regions, which is at the core of the EU discussion on how to regulate mobility.

Decentralisation with mobility versus decentralisation without mobility. Compared to a decentralised system without mobility, allowing mobility without any form of transfers generates a 'race to the bottom' with lower quality in both regions. This arises because the high-skill region has a lower marginal benefit from quality: higher quality attracts patients from the low-skill region, but does not generate any revenues. The low-skill region also has poor incentives to increase quality: lower quality shifts more patients to the high-skill region which reduces costs. An important implication is therefore that allowing mobility within the EU without any form of transfer system is undesirable.

The comparison leads to different conclusions if a system of transfers is in place. Suppose that the low-skill region pays a price equal to the marginal cost for every patient treated by the high-skill region. In this case, decentralisation with mobility can generate a (weak) Pareto improvement compared to decentralisation without mobility. The high-skill region is indifferent but the low-skill region is better off. The high-skill region is indifferent because the marginal cost of treating

the patients is exactly compensated by the price. The low-skill region benefits because patients who move to the high-skill region receive higher quality which in turn reduces the incentive of the low-skill region to provide quality. This result implies that within the EU a price system can be introduced which makes every country better off: countries that import patients can be compensated by an adequate price and countries that export patients can benefit from the higher quality.

A transfer system with a price equal to the marginal cost is generally not optimal. The optimal price which maximises the sum of the regions' welfare is strictly above the marginal cost but does not generate a Pareto improvement. The high-skill region now strictly gains thanks to the positive revenues generated by mobility and the low-skill region loses due to the higher payments to the high-skill region. Introducing such a transfer system may then be faced by stronger opposition. Within the EU this result implies that although the optimal price should be set strictly above the marginal cost to further encourage the high-skill regions to increase quality, this may be faced by the opposition of low-skill regions.

The optimal price which maximises total welfare under a decentralised solution generates nevertheless a lower welfare than under a centralised solution (which coincides with the first best). We show that a more sophisticated transfer system which entails a price paid *to* the high-skill region which is different from the price paid *by* the low-skill region does not lead to any further welfare improvements. This result holds regardless of how the extra tax bill (due to different prices) is shared between the two regions, and it arises because a marginal increase in the price paid to the high-skill region or in the price paid by the low-skill region leads to a higher quality in *both* regions. One implication of this result is that there is no need for the EU to develop a complex payment system where the price paid by exporting countries is different from the importing ones, with price differences financed through the EU budget. Instead a system with only one price is sufficient to maximise welfare.

We believe our paper is the first to study the impact of patient mobility on the provision and financing of health care across jurisdictions (regions or countries).[5] In the health economics literature there is a vast amount of papers studying the relationship between competition between health

[5] There is a paper by Petretto (2000) that looks at regionalisation of a National Health Service. It provides conditions for establishing whether devolution for health care expenditure is desirable. Variations in health expenditure will depend on its marginal benefit and the marginal cost of public funds, including higher or lower transfers originating from mobility. However, this paper has no explicit spatial dimension and it is not concerned with the quality of care. It is thus very different from ours.

care providers and their quality incentives.[6] A main lesson from this literature is that competition increases health care quality if prices are fixed (above marginal costs) and providers are profit-maximisers. However, if providers are altruistic (i.e., care about their patients), then the relationship between competition and quality is generally ambiguous (Brekke, Siciliani and Straume, 2011). The same result applies when providers also compete in prices, since then more competition depresses the profit margin of the providers, which reduces the incentive for investing in quality.

Our paper has clear parallels to this literature, since we allow patients to demand care by a competing health care provider in another region. We could also reinterpret the decentralised (centralised) solution as the competition (monopoly) solution. The question is then whether we simply can transfer the results from the previous literature to an interjurisdictional setting.[7] Our analysis and results show that the answer is no. The quality incentives are determined by a trade-off between the marginal benefit of higher quality and the marginal cost (which is the tax rate) to the patient. We show that this trade-off depends critically on the transfer payments applying to cross-border patient flows, which in turn determine the regional welfare effects of patient mobility. Thus, our analysis provides novel insights into the provision of health care quality in an interjurisdictional setting.

Our paper also relates to the economic literature on fiscal federalism,[8] in particular the part of this literature concerned with cross-border shopping. The seminal work by Kanbur and Keen (1993) provides a Hotelling model with two countries that differ in size (i.e., population density), where consumers either buy the (private) product in their home country or travel to the neighbouring country if the tax rate is significantly low. There is free entry of firms, implying a firm at every consumers' 'doorstep'. Assuming that governments are Leviathans, they show that the Nash equilibrium implies that the small country sets a lower tax rate, inducing cross-border shopping from residents in the large country. They also show that tax competition is harmful for both countries, in particular, when the difference in size is large, implying a scope for tax coordination policies.[9]

Cremer and Gahvari (2000), who study tax evasion and fiscal competition, modify the Kanbur-Keen model by introducing a public good that is financed through taxation on the private good.

[6] See Gaynor (2006) for an excellent review of the literature on competition and quality in health care markets.

[7] There is a paper by Aiura and Sanjo (2010) that uses a Hotelling model with two regions that differ in their population density to study incentives for health care quality. While this paper shares some similarities in the demand structure, the focus is very different as they study the impact of privatisation of local public hospitals.

[8] For excellent reviews of the literature on fiscal federalism, see Oates (1999, 2001).

[9] Similar results are derived by Trandel (1994), who assumes different population densities, Wang (1999), who analyses the Stackelberg equilibrium, and Nielsen (2001), who assumes a transport cost on the commodities.

They also assume that governments maximise welfare rather than being Leviathans. In this sense the paper by Cremer and Gahvari is closer to ours. However, our paper differs from the literature on cross-border shopping despite some similarities. First, in our model cross-border shopping is motivated by differences in the quality of – rather than the tax on – the good. Taxation in our model is based on the location of the consumer, not on the location of the product.[10] Second, we assume the private product to be publicly funded (through income taxation), implying an explicit link between the tax rate and the provision of the private good. Thus, the incentives for increasing taxes in our model are very different from those in Kanbur and Keen (1993), but in line with Cremer and Gahvari (2000) if we ignore the possibility of tax evasion. Finally, we do not assume free entry of firms, but rather assume that the good is not just publicly funded but also publicly provided. Considering health care markets, we believe it is appropriate to restrict attention to a limited number of firms (hospitals or physicians) rather than assuming that every consumer has a firm at its doorstep.

The rest of the paper is organised as follows. In Section 2 we present our basic modelling framework. In Sections 3 and 4 we derive the first best and the centralised solutions, respectively. In Section 5, we analyse the decentralised solution without patient mobility (Section 5.1) and with patient mobility under different payment systems (Section 5.2). Finally, in Section 6 we present some policy implications and concluding remarks.

2 Model

Consider a market for health care where consumers (patients) are uniformly distributed on a line $L = [0, 1]$. The market consists of two different regions, which can be interpreted either as two neighbouring countries or as two neighbouring regions within the same country. We will henceforth refer to the two regions as Region 1 and Region 2. Consumers located on the line segment $L_1 = [0, \frac{1}{2}]$ belong to Region 1 while the remaining consumers, located on the line segment $L_2 = [\frac{1}{2}, 1]$, belong to Region 2. The market is served by two health care providers (hospitals) which are located at the endpoints of L; thus, the provider owned by Region 1 is located at 0 while the provider owned by Region 2 is located at 1. Each patient demands one unit of health care (one treatment). We assume

[10] One can interpret this as an optimal (commodity) tax adjustment at the border; i.e., you are free to purchase the good in any country you like, but you will need to pay the home country commodity tax rate when 'importing' the good.

that health care provision is publicly funded through general income taxation and is free at the point of consumption. The utility of a patient who is located at $x_i \in L_i$ and is treated by the provider in Region j, located at z_j, is given by

$$U(x_i, z_j) = \begin{cases} y(1-\tau) + v + \beta q_j - t|x_i - z_j| & if \quad i = j \\ y(1-\tau) + v + \beta q_j - t|x_i - z_j| - f & if \quad i \neq j \end{cases}, \tag{1}$$

where $v > 0$ is the patient's gross utility of being treated, $q_j \geq \underline{q}$ is the quality offered by the provider in Region j (with $\beta > 0$ measuring the marginal utility of quality), t is marginal travelling cost, y is gross individual income and $\tau > 0$ is a proportional tax rate.[11] The lower bound $\underline{q}$ represents the lowest possible quality the providers can offer without being charged with malpractice and is, for simplicity, normalised to 0. In addition to variable travelling costs, patients also face a fixed cost $f > 0$ of travelling outside their own region for treatment. We assume that there are two types of patients: a fraction $1 - \lambda$ of the patients have a prohibitively high value of f and will always seek treatment from their local provider, while the remaining fraction λ have a low value of f and will (if allowed) travel to the neighbouring region if the quality of the treatment offered there is sufficiently high. For simplicity, we set $f = 0$ for the latter type of patients and assume that the fraction λ is constant at each point in L. Thus, we can interpret λ as an exogenous measure of the degree of interjurisdictional patient mobility. The total patient mass is normalised to 1.

If Region i faces a demand for D_i treatments, the cost of providing these treatments with a quality q_i is given by

$$C_i = cD_i + G(\theta_i, q_i), \tag{2}$$

where $c > 0$ is the marginal cost of treatment (for a given quality) and θ_i is a positive parameter that reflects the cost of quality provision, where $G(\theta_i, q) > (<) G(\theta_j, q)$ and $G_q(\theta_i, q) > (<) G_q(\theta_j, q)$ for all $q \geq 0$, if $\theta_i > (<) \theta_j$.[12] While the marginal treatment cost is assumed to be constant and equal across the two regions, we assume that Region 1 has a superior technology for providing health care quality; i.e., $\theta_1 < \theta_2$. We will therefore intermittently refer to Region 1 and Region 2 as the high-skill

[11] We may also think of τ as the social insurance contribution set by the government.

[12] For simplicity, we assume that the marginal cost of quality provision is independent of treatment volume, implying that quality is a public good for the patients of a hospital. This is a widely used assumption in the theoretical literature on quality competition between health care providers (see, e.g., Lyon, 1999; Barros and Martinez-Giralt, 2002; Gravelle and Sivey, 2010).

and low-skill regions, respectively. Several of our results in the subsequent analysis will be derived using the following quadratic form: $G(\theta_i, q_i) = \frac{\theta_i}{2} q_i^2$.

3 The first-best solution

As a benchmark for comparison, we start out by considering the first-best solution, where a utilitarian supraregional policy maker chooses the quality of each provider and also decides which patients are treated by which provider. Thus, the first-best outcome is given by the solution to the following problem:

$$\begin{aligned}\max_{x,q_1,q_2} W &= y(1-\tau) + \lambda\left(\int_0^x (v+\beta q_1 - ts)\,ds + \int_x^1 (v+\beta q_2 - t(1-s))\,ds\right) \\ &\quad + (1-\lambda)\left(\int_0^{\frac{1}{2}} (v+\beta q_1 - ts)\,ds + \int_{\frac{1}{2}}^1 (v+\beta q_2 - t(1-s))\,ds\right) \end{aligned} \tag{3}$$

subject to the budget constraint

$$\tau y = c + G(\theta_1, q_1) + G(\theta_2, q_2). \tag{4}$$

Substituting the budget constraint into the objective function and maximising, yields the following first-order conditions:

$$\left(q_1^{fb}\right) : \frac{\beta}{2}\left[1 + \lambda\left(2x^{fb} - 1\right)\right] = G_{q_1}\left(\theta_1, q_1^{fb}\right), \tag{5}$$

$$\left(q_2^{fb}\right) : \frac{\beta}{2}\left[1 - \lambda\left(2x^{fb} - 1\right)\right] = G_{q_2}\left(\theta_2, q_2^{fb}\right), \tag{6}$$

where

$$x^{fb} = \frac{1}{2}\left(1 + \frac{\beta}{t}\left(q_1^{fb} - q_2^{fb}\right)\right). \tag{7}$$

By substituting for x^{fb}, the first-order conditions for first-best quality provision can be written as

$$\left(q_1^{fb}\right) : \frac{\beta}{2}\left[1 + \frac{\lambda\beta}{t}\left(q_1^{fb} - q_2^{fb}\right)\right] = G_{q_1}\left(\theta_1, q_1^{fb}\right), \tag{8}$$

$$\left(q_2^{fb}\right) : \frac{\beta}{2}\left[1 - \frac{\lambda\beta}{t}\left(q_1^{fb} - q_2^{fb}\right)\right] = G_{q_2}\left(\theta_2, q_2^{fb}\right). \tag{9}$$

In each region, quality of health care should be provided until the point where the marginal benefit is equal to the marginal cost. Since $G_{q_1} < G_{q_2}$ for $q_1 = q_2$, the first-best quality is higher in Region 1 than in Region 2, which implies that a higher number of patients are treated in Region 1 in the first-best solution. The differences in quality levels and treatment volumes increase with the degree of patient mobility (λ). With quadratic quality costs, the first-best outcome is explicitly given by[13]

$$x^{fb} = \frac{1}{2} + \frac{\beta^2 (\theta_2 - \theta_1)}{2\left(2t\theta_1\theta_2 - \lambda\beta^2 (\theta_1 + \theta_2)\right)}, \tag{10}$$

$$q_1^{fb} = \frac{\beta\left(t\theta_2 - \lambda\beta^2\right)}{2t\theta_1\theta_2 - \lambda\beta^2 (\theta_1 + \theta_2)}, \tag{11}$$

$$q_2^{fb} = \frac{\beta\left(t\theta_1 - \lambda\beta^2\right)}{2t\theta_1\theta_2 - \lambda\beta^2 (\theta_1 + \theta_2)}, \tag{12}$$

which implies that the interregional patient flow (from Region 2 to Region 1) in the first-best outcome is given by $\lambda\left(x^{fb} - \frac{1}{2}\right) = \frac{\lambda\beta^2(\theta_2 - \theta_1)}{2\left(2t\theta_1\theta_2 - \lambda\beta^2(\theta_1 + \theta_2)\right)} > 0$.

4 The centralised solution

Now suppose that the two regions belong to the same health care jurisdiction, so that the quality of health care in each region is decided by a utilitarian central policy maker as in the previous section, but patients are free to choose their preferred provider (instead of being allocated by the central policy maker). Since patients do not pay for health care directly, the individual (among the mobile patients) who is indifferent between the provider in Region 1 and the provider in Region 2 is located at $\widehat{x}$, implicitly given by

$$y(1-\tau) + v + \beta q_i - t\widehat{x} = y(1-\tau) + v + \beta q_j - t(1-\widehat{x}),$$

which yields

$$\widehat{x} = \frac{1}{2}\left(1 + \frac{\beta}{t}(q_1 - q_2)\right). \tag{13}$$

[13] The second-order conditions are satisfied if the matrix $\begin{bmatrix} \frac{\partial^2 W}{\partial x^2} & \frac{\partial^2 W}{\partial q_1 \partial x} & \frac{\partial^2 W}{\partial q_2 \partial x} \\ \frac{\partial^2 W}{\partial q_1 \partial x} & \frac{\partial^2 W}{\partial q_1^2} & \frac{\partial^2 W}{\partial q_1 \partial q_2} \\ \frac{\partial^2 W}{\partial q_2 \partial x} & \frac{\partial^2 W}{\partial q_1 \partial q_2} & \frac{\partial^2 W}{\partial q_2^2} \end{bmatrix} = \begin{bmatrix} -2t\lambda & \beta\lambda & -\beta\lambda \\ \beta\lambda & -\theta_1 & 0 \\ -\beta\lambda & 0 & -\theta_2 \end{bmatrix}$ is negative definite, which requires $2t\theta_i > (\lambda\beta)^2$, $i = 1, 2$, and $2t\theta_1\theta_2 > \lambda\beta^2 (\theta_1 + \theta_2)$.

The optimisation problem of the policy maker is now

$$\begin{aligned}\max_{q_1,q_2} W &= y(1-\tau)+\lambda\left(\int_0^{\widehat{x}}(v+\beta q_1-ts)\,ds+\int_{\widehat{x}}^1(v+q_2-t(1-s))\,ds\right)\\&\quad+(1-\lambda)\left(\int_0^{\frac{1}{2}}(v+\beta q_1-ts)\,ds+\int_{\frac{1}{2}}^1(v+\beta q_2-t(1-s))\,ds\right)\end{aligned}\tag{14}$$

subject to (4). Let the optimal quality levels be denoted by q_i^c, $i=1,2$. It is straightforward to show that the first-order conditions for this problem coincide with the ones that secure the first-best outcome, i.e., (8)-(9), implying $q_i^c=q_i^{fb}$, $i=1,2$.[14] By comparing (7) and (13), we see that $q_i^c=q_i^{fb}$ also implies $\widehat{x}(q_1^c,q_2^c)=x^{fb}$. Thus, the centralised solution also achieves the first-best allocation of treated patients across the two regions.

Proposition 1 *The optimal quality and number patients treated in each region under the centralised solution coincide with the first-best outcome, implying higher quality in the high-skill than the low-skill region and (some) patients travelling from the low-skill to the high-skill region.*

Under the assumption of a uniform tax rate τ, implying that the tax bill is split evenly between tax payers in the two regions, regional welfare under the centralised solution can be written as

$$W_1^c=\frac{1}{2}y\left(1-\tau^{fb}\right)+\frac{1}{2}\left(v+\beta q_1^{fb}\right)-\frac{t}{8},\tag{15}$$

$$W_2^c=W_1^c-\beta\left(q_1^{fb}-q_2^{fb}\right)\left(\frac{1}{2}-\lambda\left(x^{fb}-\frac{1}{2}\right)\right)-\lambda t\left(x^{fb}-\frac{1}{2}\right)^2.\tag{16}$$

where $\tau^{fb}=[c+G(\theta_1,q_1^{fb})+G(\theta_2,q_2^{fb})]/y$. Thus, welfare is higher in the region that provides the higher level of quality (i.e., Region 1). There are two sources of this regional welfare difference: first, patients who are not treated in Region 1 suffer a utility loss from the lower quality level in Region 2,

[14] The second-order conditions are

$$\frac{\lambda\beta^2}{2t}-G_{q_1q_1}(\theta_1,q_1)<0,$$
$$\frac{\lambda\beta^2}{2t}-G_{q_2q_2}(\theta_2,q_2)<0$$

and

$$2t\left[G_{q_1q_1}(\theta_1,q_1)\,G_{q_2q_2}(\theta_2,q_2)\right]-\lambda\beta^2\left[G_{q_1q_1}(\theta_1,q_1)+G_{q_2q_2}(\theta_2,q_2)\right]>0.$$

and, second, patients who travel from Region 2 to Region 1 to enjoy the higher quality level still suffer a utility loss due to higher travelling costs. These two welfare losses are captured by, respectively, the second and third terms on the right-hand side of (16).

5 Decentralised health care provision

Suppose that the two regions belong to different jurisdictions. In each region, the optimal quality of health care is chosen to maximise the utility of patients living in that region (regardless of where they are treated), and the cost of health care provision in Region i is financed by a proportional income tax τ_i levied on the region's tax payers. We will compare two environments where interregional patient mobility is allowed or not, starting with the latter case.

5.1 No patient mobility across jurisdictions

If patients are not allowed to seek treatment in another region, the optimisation problem of the policy maker in Region i is given by

$$\max_{q_i} W_i = \frac{y}{2}(1-\tau_i) + \int_0^{\frac{1}{2}} (v + \beta q_i - ts)\, ds, \tag{17}$$

subject to

$$\frac{\tau_i y}{2} = \frac{c}{2} + G(\theta_i, q_i). \tag{18}$$

The first-order condition for optimal quality provision under no mobility, denoted q_i^n, is

$$\frac{\beta}{2} = G_{q_i}(\theta_i, q_i^n). \tag{19}$$

With decentralised health care provision, the quality of health care is still higher in Region 1 than in Region 2, due to the superior health care technology in the former region. By comparing (19) and (8)-(9), it is straightforward to verify that $q_1^n < q_1^{fb}$ and $q_2^n > q_2^{fb}$. Keeping in mind that $x^{fb} > \frac{1}{2}$, we make the following conclusion:

Proposition 2 *Compared with the first best solution, decentralisation without patient mobility is sub-optimal. In the high-skill (low-skill) region too few (many) patients are treated and the quality*

provided is too low (high).

Decentralisation without mobility is sub-optimal: the potentially mobile patients residing at the border between Region 1 and Region 2 would be willing to travel to Region 1 to obtain higher quality but are not allowed to do so, which in turn generates a welfare loss. In the absence of interregional patient mobility, the potential gains from the technological advantage of Region 1 are not fully exploited. In terms of aggregate utility across the two regions, it would have been more efficient to increase the quality difference even further and let the (mobile) patients in Region 2 who are located on the line segment $[\frac{1}{2}, x^{fb}]$ travel to Region 1 for treatment. An implication of this inefficiency is that total health care expenditures are too high in Region 2 and too low in Region 1.

We know from Proposition 1 that the first-best outcome can be implemented with centralised decisions on health care quality and free patient choice. However, even though total welfare across the two regions would be higher in a centralised solution, it is not necessarily the case that both regions would individually benefit from centralised policy making with interregional patient mobility. Regional welfare in the decentralised solution without mobility is given by

$$W_i^n = \frac{y}{2} + \frac{1}{2}\left(v + \beta q_i^n - \frac{t}{4}\right) - \frac{c}{2} - G\left(\theta_i, q_i^n\right), \quad i = 1, 2. \tag{20}$$

Whether Region 2 is better or worse off under decentralisation depends on the sign of the following expression:

$$W_2^n - W_2^c = \frac{1}{2}\beta\left(q_2^n - q_2^{fb}\right) - \frac{\lambda\beta^2}{4t}\left(q_1^{fb} - q_2^{fb}\right)^2 - G\left(\theta_2, q_2^n\right) + \frac{\left(G\left(\theta_1, q_1^{fb}\right) + G\left(\theta_2, q_2^{fb}\right)\right)}{2}. \tag{21}$$

In a decentralised solution without patient mobility, immobile patients (with prohibitively high f) and potentially mobile patients (with $f = 0$) who are located on $[x^{fb}, 1]$ enjoy a higher quality of health care than they would have in the centralised solution. On the other hand, potentially mobile patients located on $[\frac{1}{2}, x^{fb}]$ are deprived of access to higher-quality health care in Region 1 in the absence of patient mobility (since $q_2^n < q_1^{fb}$). These two welfare effects are represented by the first two terms in (21). In addition, the tax burden of residents in the two regions is generally different in the two solutions, as shown by the final two terms in (21). In the decentralised solution the cost of health care provision in Region 2 is higher, but, on the other hand, the residents of Region 2 do

not need to take part in financing the higher health care costs of Region 1.

We can derive unambiguous regional welfare effects with quadratic quality costs. In this case, equilibrium qualities in the decentralised solution without patient mobility are given by

$$q_i^n = \frac{\beta}{2\theta_i}, \quad i = 1, 2. \tag{22}$$

Using (15)-(16) and (20), Region 2 is better off in the decentralised regime if

$$W_2^n - W_2^c = \beta^2 (\theta_2 - \theta_1) \left(\frac{\beta^2 \lambda \left(2t\theta_1\theta_2 - \beta^2 \lambda (\theta_1 + \theta_2) \right) + 2t\theta_2^2 \left(t\theta_1 - \lambda\beta^2 \right)}{8\theta_2 \left(2t\theta_1\theta_2 - \beta^2 \lambda (\theta_1 + \theta_2) \right)^2} \right) > 0, \tag{23}$$

which is true for all valid parameter configurations.[15] Since centralised policy making with mobility implements the first-best outcome, but Region 2 prefers decentralised policy making without mobility, then welfare in Region 1 must necessarily be higher in the centralised solution:[16]

Proposition 3 *Compared with decentralisation without patient mobility, the high-skill (low-skill) region is better (worse) off under centralised policy-making with interregional patient mobility.*

Thus, even if centralised policy making implements the first-best outcome, the low-skill region (Region 2) would not be willing to transfer authority to a central policy maker unless there is a system of compensation (e.g., an interregional income transfer policy) in place.

Notice that $W_2^n - W_2^c > 0$ also for $\lambda = 0$. Thus, even if allowing for patient mobility does not actually lead to any outflow of patients from Region 2 (which implies $q_i^n = q_i^{fb}$), this region is still better off under decentralised policy making. The reason is that, when health care is financed by uniform income taxation, tax payers in Region 2 must contribute to financing the higher health care expenditures in Region 1 in the centralised solution.[17] If allowing for patient mobility leads to an outflow of patients from Region 2 in equilibrium (i.e., $\lambda > 0$), there are two additional welfare effects of centralisation for Region 2: (i) lower utility for the patients who are treated in Region 2 (since $q_2^{fb} < q_2^n$), and (ii) higher utility for the patients who travel to Region 1 for treatment (since $q_1^{fb} > q_2^n$). We can show that the first effect dominates, implying that higher patient mobility

[15] Notice that the numerator in (23) is positive due to the second-order conditions and $q_2^c \geq 0$.

[16] Since $W_1^c + W_2^c > W_1^n + W_2^n$ and $W_2^n > W_2^c$ it follows that $W_1^c > W_1^n$.

[17] Since the marginal cost of quality provision is lower in Region 1 than in Region 2, the corresponding higher quality level in Region 1 implies that the total cost of qualty provision is also higher in this region. Although this result holds for quadratic quality costs, it does not generalise to any convex cost function.

increases the welfare loss of centralisation for the low-skill region, if the degree of patient mobility (as measured by λ) is sufficiently low to begin with.[18]

5.2 Interjurisdictional patient mobility

This section derives quality choices and (regional and total) welfare when patients' mobility is allowed under four plausible scenarios: (i) no transfer system is in place; (ii) the transfer system sets the price equal to the marginal cost; (iii) the transfer price is determined to maximise total welfare (defined as the sum of regions' utility); (iv) the price paid by the exporting region is different from the price received by the importing region. We then address the following key question: compared to decentralisation without mobility, how does interjurisdictional patient mobility affect quality choices and regional welfare? We also compare qualities under decentralisation and mobility with the first-best ones.

Suppose that individuals are free to choose the health care provider they prefer, regardless of whether the provider and the patient belong to the same health care jurisdiction. The two policy makers are assumed to choose the quality of health care in their respective regions non-cooperatively. Since Region 1 has a superior technology for providing health care quality, there will be an outflow of patients from the low-skill to the high-skill region in equilibrium. The size of this patient flow is determined by the share of mobile patients (λ) and the location of the indifferent patient among these ($\widehat{x}$). We assume that the health care provider in Region 1 cannot turn down patients who travel from Region 2 to obtain treatment. How is the health care to these patients paid for? Suppose that Region 2 pays a transfer to Region 1. We assume that this transfer takes the form of a price p for each of its own residents who are treated in Region 1. We will first derive the Nash equilibrium for any given p, and subsequently explore the four plausible pricing rules outlined above.

Anticipating that $\widehat{x} > \frac{1}{2}$ in equilibrium, the optimisation problem of the policy maker in Region 1 is

$$\max_{q_1} W_1 = \frac{y}{2}(1-\tau_1) + \int_0^{\frac{1}{2}} (v + \beta q_1 - ts)\, ds, \tag{24}$$

subject to

$$\frac{\tau_1 y}{2} = \frac{c}{2} - (p-c)\,\lambda\left(\widehat{x} - \frac{1}{2}\right) + G(\theta_1, q_1), \tag{25}$$

[18] $\frac{\partial\left(W_2^n - W_2^c\right)}{\partial\lambda} = \frac{t\beta^4(\theta_2-\theta_1)\left[4t\theta_1^2\theta_2 - \beta^2\lambda(\theta_1+\theta_2)^2\right]}{4\left(2t\theta_1\theta_2 - \beta^2\lambda(\theta_1+\theta_2)\right)^3}$. This expression is always positive if λ is sufficiently low.

while the optimisation problem of Region 2 is

$$\begin{aligned}\max_{q_2} W_2 &= \frac{y}{2}(1-\tau_2) + \lambda \int_{\frac{1}{2}}^{\widehat{x}} (v + \beta q_1 - ts)\, ds \\ &\quad + (1-\lambda) \int_{\frac{1}{2}}^{\widehat{x}} (v + \beta q_2 - t(1-s))\, ds + \int_{\widehat{x}}^{1} (v + \beta q_2 - t(1-s))\, ds, \end{aligned} \tag{26}$$

subject to

$$\frac{\tau_2 y}{2} = \frac{c}{2} + (p-c)\lambda\left(\widehat{x} - \frac{1}{2}\right) + G(\theta_2, q_2). \tag{27}$$

Notice that the second term on the right-hand side of (25) represents the net revenue for Region 1 of treating patients from the neighbouring region, while the second term on the right-hand side of (27) represents the corresponding net cost for Region 2.

The first-order conditions that define the Nash equilibrium under decentralisation and patient mobility, with equilibrium qualities denoted by q_i^m, $i = 1, 2$, are given by[19]

$$(q_1^m) : \frac{\beta}{2}\left(1 + \frac{\lambda}{t}(p-c)\right) = G_{q_1}(\theta_1, q_1^m), \tag{28}$$

$$(q_2^m) : \frac{\beta}{2}\left(1 + \frac{\lambda}{t}(p - c - \beta(q_1^m - q_2^m))\right) = G_{q_2}(\theta_2, q_2^m). \tag{29}$$

In each region, the level of health care quality is chosen such that the marginal utility for the region's residents plus the marginal net revenue from interregional patient flows are equal to the marginal cost of quality provision. In equilibrium, the health care quality is always higher in Region 1 than in Region 2, and the welfare in each region is given by

$$W_1^m = \frac{y}{2} + \frac{1}{2}(v + \beta q_1^m) - \frac{t}{8} - \frac{1}{2}\left(c - \frac{\lambda\beta}{t}(q_1^m - q_2^m)(p-c)\right) - G(\theta_1, q_1^m) \tag{30}$$

and

$$W_2^m = \frac{y}{2} + \frac{1}{2}(v + \beta q_2^m) - \frac{t}{8} + \lambda\frac{\beta^2}{4t}(q_1^m - q_2^m)^2 - \frac{1}{2}\left(c + \frac{\lambda\beta}{t}(q_1^m - q_2^m)(p-c)\right) - G(\theta_2, q_2^m). \tag{31}$$

Before considering different rules for choosing p, let us first see how equilibrium qualities depend on the level of p. By totally differentiating (28)-(29) and applying Cramer's rule, we can show that

[19] The second-order conditions are $-G_{q_1 q_1} < 0$ and $\frac{\lambda\beta^2}{2t} - G_{q_2 q_2} < 0$.

the equilibrium quality responses to a marginal increase in the transfer payment p are given by

$$\frac{\partial q_1^m}{\partial p} = \frac{\lambda\beta}{2tG_{q_1q_1}\left(\theta_1, q_1^m\right)} > 0, \tag{32}$$

$$\frac{\partial q_2^m}{\partial p} = \frac{\lambda\beta\left(G_{q_1q_1}\left(\theta_1, q_1^m\right) - \frac{\lambda\beta^2}{2t}\right)}{2tG_{q_1q_1}\left(\theta_1, q_1^m\right)\left(G_{q_2q_2}\left(\theta_2, q_2^m\right) - \frac{\lambda\beta^2}{2t}\right)} > 0. \tag{33}$$

Notice that the positive sign of $\partial q_2^m/\partial p$ is determined by invoking the second-order conditions of the centralised optimisation problem. The intuition for the positive relationship between the transfer payment and equilibrium qualities in the two regions is reasonably straightforward. An increase in p makes it more profitable for Region 1 to treat patients from Region 2, while it becomes more costly for Region 2 to pay for the treatment of these patients. All else equal, this gives the policy maker in Region 1 incentives to provide higher quality in order to attract more patients from the neighbouring region, while the policy maker in Region 2 has an incentive to increase quality in order to dampen the outflow of patients. In other words, a higher transfer payment intensifies quality competition between the two regions. This effect is stronger the higher the share of mobile patients (λ) and the lower the travelling costs (t). With quadratic quality costs, equilibrium qualities in the two regions are given by

$$q_1^m = \frac{\beta\left(t + \lambda\left(p - c\right)\right)}{2t\theta_1}, \tag{34}$$

$$q_2^m = \frac{\beta\left(2t\theta_1 - \lambda\beta^2\right)\left(\lambda\left(p - c\right) + t\right)}{2t\theta_1\left(2t\theta_2 - \lambda\beta^2\right)}. \tag{35}$$

In the subsequent analysis, we investigate the four pricing rules mentioned at the beginning of this section.

5.2.1 No transfer payment

Suppose that there is no system of transfer payment in place; i.e., $p = 0$. The Nash equilibrium is then characterised by

$$\left(q_1^m\right) : \frac{\beta}{2}\left(1 - \frac{\lambda c}{t}\right) = G_{q_1}\left(\theta_1, q_1^m\right), \tag{36}$$

$$\left(q_2^m\right) : \frac{\beta}{2}\left(1 - \frac{\lambda}{t}\left(c + \beta\left(q_1^m - q_2^m\right)\right)\right) = G_{q_2}\left(\theta_2, q_2^m\right). \tag{37}$$

Comparing with the case of decentralisation without mobility, (19), we see that patient mobility leads to lower health care quality in both regions: $q_1^m < q_1^n$ and $q_2^m < q_2^n$. Since Region 1 is not compensated for the treatment of patients from Region 2, the policy maker in Region 1 has an incentive to reduce the quality in order to dampen the inflow of such patients. At the same time, the policy maker in Region 2 has an incentive to stimulate patient outflow, by reducing quality, in order to pass some of the region's health care expenditures on to the tax payers of Region 1. In other words, Region 2 has an incentive to *free ride* on the high-skilled Region 1's quality investments. Thus, allowing for interjurisdictional patient mobility without transfer payments leads to a 'race to the bottom' in terms of health care quality. In fact, even if a transfer payment scheme is in place, a race-to-the-bottom effect is present – albeit in a milder form – for any price below marginal cost ($p < c$).

Compared with the first-best outcome, there is clearly underprovision of quality in Region 1, while the quality in Region 2 might be higher or lower than in the first-best solution. With quadratic quality costs, quality is underprovided also in Region 2 if

$$q_2^{fb} - q_2^m = \lambda\beta \frac{c\left(2t\theta_1 - \lambda\beta^2\right)\left(2t\theta_1\theta_2 - \lambda\beta^2\left(\theta_1+\theta_2\right)\right) - \lambda t\beta^4\left(\theta_2-\theta_1\right)}{2t\theta_1\left(2t\theta_2-\lambda\beta^2\right)\left(2t\theta_1\theta_2-\lambda\beta^2\left(\theta_1+\theta_2\right)\right)} > 0. \tag{38}$$

This condition holds if the marginal treatment costs (c) is sufficiently high relative to the marginal willingness-to-pay for quality (β). The underprovision of quality in Region 1 also means that the interregional patient flow is too small.[20]

How does patient mobility affect regional welfare? For Region 1, the welfare effect of allowing interjurisdictional patient mobility is given by

$$W_1^m - W_1^n = \frac{\beta}{2}\left(q_1^m - q_1^n\right) - \left(G\left(\theta_1, q_1^m\right) - G\left(\theta_1, q_1^n\right)\right) - \frac{\lambda\beta}{2t}c\left(q_1^m - q_2^m\right) < 0. \tag{39}$$

Welfare in Region 1 is lower with mobility for two reasons. First, the quality of health care goes

[20] The indifferent patient (among the mobile ones) is located at

$$\widehat{x}\left(q_1^m, q_2^m\right) = \frac{t\left(2t\theta_1\theta_2 + \beta^2\left(\theta_2-\theta_1\right)\right) - \lambda\beta^2\left(t\theta_1 + c\left(\theta_2-\theta_1\right)\right)}{2t\theta_1\left(2t\theta_2-\lambda\beta^2\right)}.$$

The comparison with the first-best solution is given by

$$\widehat{x}\left(q_1^m, q_2^m\right) - x^{fb} = -\lambda\beta^2\left(\theta_2-\theta_1\right)\frac{c\left(2t\theta_1\theta_2 - \lambda\beta^2\left(\theta_1+\theta_2\right)\right) + t\beta^2\theta_2}{2t\theta_1\left(2t\theta_2-\beta^2\lambda\right)\left(2t\theta_1\theta_2-\lambda\beta^2\left(\theta_1+\theta_2\right)\right)} < 0.$$

down. The ensuing welfare loss of this drop in quality is given by the sum of the two first terms in (39).[21] Second, tax payers in Region 1 must pay for the treatment of patients travelling from Region 2, the cost of which is given by the third term.

For Region 2, the welfare effect of patient mobility is *a priori* more ambiguous:

$$W_2^m - W_2^n = \frac{\beta}{2}\left(q_2^m - q_2^n\right) - \left(G\left(\theta_2, q_2^m\right) - G\left(\theta_2, q_2^n\right)\right) + \lambda\frac{\beta^2}{4t}\left(q_1^m - q_2^m\right)^2 + \frac{\lambda\beta}{2t}c\left(q_1^m - q_2^m\right). \quad (40)$$

Mobility has three effects on welfare in Region 2. First, quality goes down for the patients that are treated within the region. Second, since the ranking of q_1^m and q_2^n is ambiguous, the patients who take advantage of interregional mobility and seek treatment in Region 1 may enjoy higher or lower quality.[22] Third, the tax burden goes down since some of the health care expenditures are passed on to Region 1 through interregional patient mobility. The sum of the first and second effect is given by the sum of the three first terms in (40), while the third effect is given by the last term. Notice that, if $\theta_1 \to \theta_2$, there is no interregional patient mobility in equilibrium and the second and third effect vanish, implying that $W_2^m - W_2^n < 0$. Due to continuity, mobility will lead to lower welfare in Region 2 also for a sufficiently small technology difference $(\theta_2 - \theta_1)$ between the regions. With quadratic quality costs, however, the ambiguity is resolved and mobility leads to a welfare reduction for all valid parameter configurations:

$$W_2^m - W_2^n = -\frac{\lambda^2\beta^2\left(t\beta^2\left(\theta_2 - \theta_1\right) + c\theta_2\left(2t\theta_1 - \beta^2\lambda\right)\right)^2}{8t^2\theta_1^2\theta_2\left(2t\theta_2 - \beta^2\lambda\right)^2} < 0. \quad (41)$$

Proposition 4 *Under decentralisation, allowing for interjurisdictional patient mobility with $p = 0$ leads to lower quality and lower welfare in both regions. Compared with the first-best outcome, quality is always underprovided in the high-skill region and is also underprovided in the low-skill region if marginal treatment costs are sufficiently high relative to the marginal utility of health care quality.*

Proposition 4 makes a clear case against allowing mobility if a transfer system is not in place.

[21] Notice that

$$\frac{\beta}{2}\left(q_1^m - q_1^n\right) - \left(G\left(\theta_1, q_1^m\right) - G\left(\theta_1, q_1^n\right)\right) < 0$$

since q_1^n maximises $\frac{\beta q_1}{2} - G\left(\theta_1, q_1\right)$.

[22] With quadratic quality costs, $q_1^m < q_2^n$ if $c > \frac{t(\theta_2 - \theta_1)}{\lambda\theta_2}$. In other words, if the technological difference between the regions is low relative to the marginal treatment cost, patient mobility leads to lower quality for *all* patients, including those patients who travel from the low-skill region to obtain health care in the high-skill region.

Under a decentralised solution with mobility, regions would have poor incentives to provide quality, which leads to low welfare in both regions.

5.2.2 Transfer payment equal to marginal cost

Suppose that the transfer payment is set equal to marginal treatment costs; i.e., $p = c$. In this case, the Nash equilibrium is characterised by

$$(q_1^m) : \frac{\beta}{2} = G_{q_1}\left(\theta_1, q_1^m\right), \tag{42}$$

$$(q_2^m) : \frac{\beta}{2}\left(1 - \frac{\lambda\beta}{t}\left(q_1^m - q_2^m\right)\right) = G_{q_2}\left(\theta_2, q_2^m\right). \tag{43}$$

For Region 1, since the transfer payment exactly covers the cost of treating patients from the other region, the incentives to provide quality are unaffected by whether interregional patient mobility is allowed or not, so that $q_1^m = q_1^n$. The policy maker in Region 2, on the other hand, has an incentive to reduce quality when interregional mobility is allowed. Since a fraction of the region's patients can obtain health care of higher quality in the neighbouring region, the marginal welfare gain of quality provision in Region 2 is lower, and regional welfare is thus maximised at a lower quality level, so that $q_2^m < q_2^n$.

Since patient mobility with $p = c$ does not affect quality incentives in Region 1, equilibrium quality in this region is underprovided relative to the first-best solution: $q_1^m < q_1^{fb}$. In Region 2, patient mobility reduces quality incentives and brings equilibrium quality closer to the first-best level in this region. However, comparing (9) and (43) we see that $q_1^m < q_1^{fb}$ implies $q_2^m > q_2^{fb}$, which means that quality is still overprovided in Region 2.

Due to unchanged incentives for quality provision, allowing for interregional patient mobility has no effect on welfare in Region 1, and $W_1^m = W_1^n$. In Region 2, the effect is indeterminate:

$$W_2^m - W_2^n = \frac{\lambda\beta^2}{4t}\left(q_1^m - q_2^m\right)^2 - \frac{\beta}{2}\left(q_2^n - q_2^m\right) + \left(G\left(\theta_2, q_2^n\right) - G\left(\theta_2, q_2^m\right)\right). \tag{44}$$

Mobility affects different types of patients differently. Some patients from Region 2 get access to higher-quality health care in Region 1 (first term), while the remaining patients in Region 2 experience a drop in the quality of the health care they are offered (second term) which implies a lower cost

of quality provision (third term). Figure 1 shows graphically how different groups of patients are affected: patients from Region 1 in area A are indifferent; patients from Region 2 in area B receive a higher quality while patients in area C receive a lower one.

[Figure 1 about here]

Once more, we proceed to assuming quadratic quality costs, in which case the equilibrium is given by

$$q_1^m = \frac{\beta}{2\theta_1}, \tag{45}$$

$$q_2^m = \frac{2t\beta\theta_1 - \lambda\beta^3}{\theta_1\left(4t\theta_2 - 2\beta^2\lambda\right)}, \tag{46}$$

and the welfare effect of mobility in Region 2 is

$$W_2^m - W_2^n = \frac{\lambda\beta^4\left(\theta_2 - \theta_1\right)^2}{8\theta_1^2\theta_2\left(2t\theta_2 - \lambda\beta^2\right)} > 0. \tag{47}$$

Thus, the gain of the patients who get access to higher-quality care in Region 1 outweighs the loss of the remaining patients who have to accept a quality degradation.

Proposition 5 *Under decentralisation, allowing for interjurisdictional patient mobility with $p = c$ has no effect on quality and welfare in the high-skill region, while quality goes down and welfare goes up in the low-skill region. Compared with the first-best outcome, quality is underprovided in the high-skill region and overprovided in the low-skill region.*

Proposition 5 suggests that under decentralisation with mobility a transfer system which sets the price equal to the marginal treatment cost can generate a (weak) Pareto improvement compared to decentralisation without mobility. The high-skill region is indifferent between mobility and no mobility because the price covers the marginal cost of treating the patients. The low-skill region benefits since patients moving to the high-skill region receive higher quality, despite the reduction of quality for the patients who do not move.

Since there are both winners and losers from mobility within the low-skill region, we can paint a more detailed picture of the welfare effects of mobility in this region by assessing the relative sizes of these two groups of patients. Notice that, among the potentially mobile patients (with $f = 0$)

in Region 2, a patient who is located to the left (right) of $\widetilde{x}$ is better (worse) off with than without mobility, where $\widetilde{x}$ is implicitly given by

$$\beta q_1^m - t\widetilde{x} - 2G\left(\theta_2, q_2^m\right) = \beta q_2^n - t\left(1 - \widetilde{x}\right) - 2G\left(\theta_2, q_2^n\right). \tag{48}$$

With quadratic quality costs, the location of the patient who is indifferent between mobility and no mobility is given by

$$\widetilde{x} = \frac{1}{2} + \frac{\left(4\theta_2\theta_1 t\left(2t\theta_2 - \beta^2\lambda\right) - \lambda^2\beta^4\left(\theta_2 - \theta_1\right)\right)\left(\theta_2 - \theta_1\right)\beta^2}{8\left(2t\theta_2 - \beta^2\lambda\right)^2 t\theta_1^2\theta_2}. \tag{49}$$

Since patients are uniformly distributed, with a mass of $\frac{1}{2}$ in each region, the winners from mobility outnumber the losers if

$$\lambda\left(\widetilde{x} - \frac{1}{2}\right) > \frac{1}{4}.$$

It is reasonably straightforward to show that a necessary (but not sufficient) condition for this inequality to hold is

$$\lambda > \frac{2t\theta_1\theta_2}{\beta^2\left(2\theta_2 - \theta_1\right)}. \tag{50}$$

This condition *does not hold* if there are sufficiently few potentially mobile patients, if travelling costs are sufficiently high, or if the marginal utility of quality is sufficiently low, and it is *less likely to hold* if the skill-difference between the two regions is sufficiently low. Thus, even if mobility leads to higher total welfare in the low-skill region, the losers from mobility will nevertheless outnumber the winners if the condition given in (50) does not hold. One possible implication of this is that the low-skill region could still block the implementation of patient mobility, if this decision is based on a process that reflects the relative number of winners and losers from the proposed policy (e.g., majority voting).

5.2.3 Optimal transfer payment

The optimal transfer payment for patients who seek treatment in a different region is a price p^* that maximises total welfare in the two regions. Comparing first-best quality provision with equilibrium quality provision under decentralisation and patient mobility, i.e., comparing (8)-(9) with (28)-(29),

we see that a price $p = c+\beta\left(q_1^{fb} - q_2^{fb}\right)$ will yield first-best quality provision in Region 1 but overprovision of quality in Region 2. Thus, the first-best outcome cannot be implemented with decentralised policy making. Since equilibrium quality provision in both regions is monotonically increasing in p, the optimal transfer payment must thus necessarily imply underprovision (overprovision) of quality in the high-skill (low-skill) region.

With quadratic quality costs, equilibrium quality provision under decentralisation and patient mobility is given by (34)-(35). In order to find the optimal transfer payment, we substitute these equilibrium quality expressions into the regional welfare functions and maximise total welfare with respect to p, yielding

$$p^* = c + \frac{t\beta^2\left(\theta_2 - \theta_1\right)}{\left(\theta_1 + \theta_2\right)\left(2t\theta_1 - \lambda\beta^2\right)}. \tag{51}$$

A first observation is that the optimal price is higher than marginal treatment costs. Thus, in order to maximise total welfare, the high-skill region must earn a strictly positive (marginal) profit from treating patients from the low-skill region. Notice also that a higher patient mobility (λ) increases the optimal price. Increased patient mobility implies that a larger number of patients in Region 2 are potentially willing to travel to Region 1 to obtain higher-quality treatment. In terms of total welfare, this increases the marginal welfare gain of quality provision in Region 1. However, since this extra welfare gain accrue to patients residing in Region 2, and the policy maker in Region 1 does not take this gain into account, a higher payment for these patients is needed in order to provide the correct incentives for quality provision in the high-skill region.

With optimal transfer payments, the quality provision in equilibrium is given by

$$q_1^m = \beta\frac{t\left(\theta_1 + \theta_2\right) - \lambda\beta^2}{\left(\theta_1 + \theta_2\right)\left(2t\theta_1 - \lambda\beta^2\right)}, \tag{52}$$

$$q_2^m = \beta\frac{t\left(\theta_1 + \theta_2\right) - \lambda\beta^2}{\left(\theta_1 + \theta_2\right)\left(2t\theta_2 - \lambda\beta^2\right)}. \tag{53}$$

The effects of patient mobility on health care provision in the two regions are given by

$$q_1^m - q_1^n = \frac{\lambda\beta^3\left(\theta_2 - \theta_1\right)}{2\theta_1\left(\theta_1 + \theta_2\right)\left(2t\theta_1 - \lambda\beta^2\right)} > 0, \tag{54}$$

$$q_2^m - q_2^n = -\frac{\lambda\beta^3\left(\theta_2 - \theta_1\right)}{2\theta_2\left(\theta_1 + \theta_2\right)\left(2t\theta_2 - \lambda\beta^2\right)} < 0. \tag{55}$$

Thus, with optimal transfer payments, allowing for interjurisdictional patient mobility will increase quality provision in the high-skill region and reduce quality provision in the low-skill region. Still, compared with the first-best outcome, there is overprovision (underprovision) in the former (latter) region.

The regional welfare effect of patient mobility turns out to be positive for Region 1:

$$W_1^m - W_1^n = \lambda\beta^4\left(\theta_2 - \theta_1\right)^2 \frac{8t^2\theta_1\left(\theta_1 + \theta_2\right) - \lambda\beta^2\left(2t\left(4\theta_1 + \theta_2\right) - \lambda\beta^2\right)}{8\theta_1\left(\theta_1 + \theta_2\right)^2\left(2t\theta_1 - \lambda\beta^2\right)^2\left(2t\theta_2 - \lambda\beta^2\right)} > 0. \tag{56}$$

Although quality is provided at a level where the marginal benefit for the patients living in Region 1 is lower than the marginal cost of quality provision, this is more than compensated for by the profit that the region makes from treating patients that travel from Region 2.[23]

For Region 2, the welfare effect of mobility is given by

$$W_2^m - W_2^n = \lambda\beta^4\left(\theta_2 - \theta_1\right)^2 \frac{\lambda\beta^2\left(2t\left(\theta_2 - 2\theta_1\right) + \lambda\beta^2\right) - 4t^2\left(\theta_2 - \theta_1\right)\left(\theta_1 + \theta_2\right)}{8\theta_2\left(\theta_1 + \theta_2\right)^2\left(2t\theta_1 - \lambda\beta^2\right)^2\left(2t\theta_2 - \lambda\beta^2\right)} < 0. \tag{57}$$

Allowing for patient mobility implies that the quality offered to the patients who are treated in Region 2 goes down. In addition, the tax payers in Region 2 have to finance part of the health care expenditures of the neighbouring region in the form of a transfer payment in excess of marginal treatment costs for the patients who seek health care outside the region. The higher quality of care enjoyed by these patients is not sufficient to fully compensate for the above mentioned welfare losses and the welfare of Region 2 is consequently reduced as a result of interjurisdictional patient mobility with optimal transfer payments.[24]

Proposition 6 *Under decentralisation, allowing for interjurisdictional patient mobility with transfer*

[23] The positive sign of $W_1^m - W_1^n$ is established by noticing that the numerator in (56) is monotonically increasing in t :

$$\frac{\partial\left(8t^2\theta_1\left(\theta_1 + \theta_2\right) - \lambda\beta^2\left(2t\left(4\theta_1 + \theta_2\right) - \lambda\beta^2\right)\right)}{\partial t} = 2\left(4\theta_1\left(2t\theta_1 - \lambda\beta^2\right) + \theta_2\left(8t\theta_1 - \lambda\beta^2\right)\right) > 0$$

Setting t at the lowest permissible value, $t = \frac{\lambda\beta^2}{2\theta_1}$, yields

$$8t^2\theta_1\left(\theta_1 + \theta_2\right) - \lambda\beta^2\left(2t\left(4\theta_1 + \theta_2\right) - \lambda\beta^2\right) = \lambda^2\beta^4\frac{\theta_2 - \theta_1}{\theta_1} > 0.$$

[24] The negative sign of $W_2^m - W_2^n$ is established by noticing that the numerator in (57) is monotonically decreasing in t :

payments that maximise total welfare leads to higher (lower) quality and welfare in the high-skill (low-skill) region. Compared with the first-best outcome, quality is underprovided in the high-skill region and overprovided in the low-skill region.

In contrast to the result obtained in Proposition 5, allowing for mobility in a decentralised system with transfer prices that maximise global welfare is not a Pareto improvement, despite generating an overall increase in welfare. It is the low-skill region which will oppose mobility: despite the increase in quality for the mobile patients, the price paid to the high-skill region is too high.

5.2.4 Transfer payments with two prices

The above analysis showed that the first-best outcome cannot be implemented under a decentralised regime with a single transfer price, even if this price is optimally chosen. In the following, we check whether we can get closer to the first-best solution by designing a transfer payment scheme with two different prices. We show that we cannot.

Suppose that Region 2 pays p_2, while Region 1 receives p_1, for each patient from Region 2 seeking treatment in Region 1. If $p_1 \neq p_2$, what Region 2 pays is not equal to what Region 1 receives. If $p_1 > p_2$, more tax revenues need to be raised; if $p_1 < p_2$, there is an extra surplus that can be distributed to tax payers. Since the number of patients travelling from Region 2 to Region 1 to seek treatment is $\lambda\left(\widehat{x}-\frac{1}{2}\right)$, the extra tax bill, in the case of $p_1 > p_2$, is $(p_1-p_2)\,\lambda\left(\widehat{x}-\frac{1}{2}\right)$. We assume that tax payers in Region 1 pay a share α of this extra bill, while tax payers in Region 2 pay the remaining share $1-\alpha$.

With this new transfer payment scheme, the policy maker in Region 1 maximises (24) subject to the new budget constraint

$$\frac{\tau_1 y}{2} = c\left(\frac{1}{2}+\left(\widehat{x}-\frac{1}{2}\right)\right) + G\left(q_1,\theta_1\right) - p_1\lambda\left(\widehat{x}-\frac{1}{2}\right) + \alpha\left(p_1-p_2\right)\lambda\left(\widehat{x}-\frac{1}{2}\right), \tag{58}$$

$$\frac{\partial\left(\lambda\beta^2\left(2t\left(\theta_2-2\theta_1\right)+\lambda\beta^2\right)-4t^2\left(\theta_2-\theta_1\right)\left(\theta_1+\theta_2\right)\right)}{\partial t} = -2\left(\theta_2\left(4t\theta_2-\lambda\beta^2\right)-2\theta_1\left(2t\theta_1-\lambda\beta^2\right)\right) < 0 \text{ for } \theta_1<\theta_2.$$

Setting t at the lowest permissible level, $t=\frac{\lambda\beta^2}{2\theta_1}$, yields

$$\lambda\beta^2\left(2t\left(\theta_2-2\theta_1\right)+\lambda\beta^2\right)-4t^2\left(\theta_2-\theta_1\right)\left(\theta_1+\theta_2\right) = -\lambda^2\theta_2\beta^4\frac{\theta_2-\theta_1}{\theta_1^2} < 0.$$

while the policy maker in Region 2 maximises (26) subject to

$$\frac{\tau_2 y}{2} = \left((1-\lambda)\left(\widehat{x}-\frac{1}{2}\right)+1-\widehat{x}\right)c+G\left(q_2,\theta_2\right)+p_2\lambda\left(\widehat{x}-\frac{1}{2}\right)+(1-\alpha)\left(p_1-p_2\right)\lambda\left(\widehat{x}-\frac{1}{2}\right). \quad (59)$$

The first-order conditions that define the Nash equilibrium are

$$\left(q_1^m\right): \frac{\beta}{2}\left(1+\frac{\lambda}{t}\left((1-\alpha)p_1+\alpha p_2-c\right)\right) = G_{q_1}\left(\theta_1, q_1^m\right), \quad (60)$$

$$\left(q_2^m\right): \frac{\beta}{2}\left(1+\frac{\lambda}{t}\left((1-\alpha)p_1+\alpha p_2-c-\beta\left(q_1^m-q_2^m\right)\right)\right) = G_{q_2}\left(\theta_2, q_2^m\right). \quad (61)$$

Comparing (28)-(29) with (60)-(61) we see that a transfer payment scheme with two different prices cannot improve upon the equilibrium outcome with a single, optimally chosen, transfer price p^*. For any pair of prices, (p_1, p_2), a uniform price $p = (1-\alpha)p_1 + \alpha p_2$ yields exactly the same outcome in terms of equilibrium qualities. This is true regardless of the value of α, i.e., regardless of how the extra tax bill is shared between the two regions.

The reason for this result is that both prices have a qualitatively similar impact on equilibrium quality in both regions; more specifically, an increase in either of the two prices leads to higher quality in both regions. An increase in p_1 has two effects on quality incentives in Region 1. It increases both the revenues from treating out-of-region patients and the extra tax revenues needed to finance the transfer payment scheme. However, since the extra tax burden is shared between the two regions, the first effect always dominates. For Region 2, an increase in p_1 has only one effect; it increases the tax burden due to a higher cost of the transfer payment scheme and therefore gives an incentive to provide higher quality in order to dampen the demand for out-of-region treatments.

An increase in p_2 has qualitatively similar effects. It becomes more costly for Region 2 to pay for out-of-region treatments, but the tax burden also goes down since the cost of the transfer payment system is reduced. However, since the gain from a lower tax burden is shared with tax payers in Region 1, the first effect dominates and equilibrium quality goes up in Region 2. The same happens in Region 1, since the indirect cost of treating out-of-region patients – in the form of higher taxes to finance the transfer payment scheme – is reduced when p_2 goes up.

Proposition 7 *Under decentralisation, a transfer payment with two prices, where the price paid by the low-skill region is different from the price received by the high-skill region, does not increase*

welfare compared to a transfer payment with one price.

6 Concluding remarks

In this paper we have studied the impact of patient mobility on health care quality, health care financing (taxation) and social welfare. Our analysis has been motivated by the new EU legislation that aims at stimulating patient mobility across EU member states, but the analysis applies also to patient mobility within countries with separate regional jurisdictions, like in Canada, Italy or Sweden, among others. To study the effects of patient mobility, we made use of a Hotelling model with two regions (countries) that differ in health care quality technology, patients that decide which region to demand medical treatment, and policy makers that decide on the health care quality and income taxation in their region. Based on this set up, we compared the decentralised solution without patient mobility (the old system within the EU) with (i) the centralised solution with patient mobility (the optimal system) and (ii) the decentralised solution with patient mobility (the new system within the EU).

We first showed that the *centralised solution with patient mobility* implements the first best. However, this case is not feasible for two reasons: first, the low-skill region gets lower welfare compared with the decentralised solution without patient mobility and is thus not willing to transfer authority to an interregional policy maker (the EU). Second, the centralised solution implies that an interregional policy maker (the EU) takes over the health care financing directly, which is a highly unlikely scenario at the moment.

We then analysed the *decentralised solution with patient mobility* under various transfer payment schemes across the regions. The 'worst-case' scenario is when there is no transfer payments, since patient mobility then leads to a 'race-to-the-bottom' effect in terms of health care quality, implying lower welfare in both regions compared with no patient mobility. Thus, in absence of transfer payments both regions will oppose patient mobility. This situation is likely to describe the current EU situation with no clear payment rules and thus very low patient mobility.

The 'feasible-case' scenario is when the transfer payment is set equal to the marginal treatment cost. In this case we have a weak Pareto improvement, where the low-skill region benefits from access to treatment in the high-skill region, whereas the welfare in the high-skill region is unchanged since

the provider is fully compensated for the extra treatment cost through the transfer payment. Notice, however, that the scope for a weak Pareto improvement is defined at regional level. Despite the low-skill region being overall better off under this pricing system, we can identify a group of winners and a group of losers: the winners belong to a subgroup of the patients who travel to the high-skill region and benefit from the higher quality (the ones living closer to the border), while the remaining patients in Region 2 will lose from allowing mobility. If the group of losers is sufficiently large, a move to a system with mobility where the price equals the marginal cost may be politically unsustainable in the low-skill region despite the potential welfare improvement.

The 'best-case' scenario is when the transfer price is set to maximise joint regional welfare. We showed that the socially optimal transfer price is higher than the marginal treatment cost, and brings the outcome closer to (but not at) the first best. The high-skill region benefits since health care quality and welfare is higher, but the low-skill region loses because of the higher taxes needed to finance the high-quality care to residents seeking care in the high-skill region. We also showed that a more complex payment system, with different (optimal) prices to the two regions cannot solve the problem. Thus, optimal transfer pricing is not a feasible scenario, since the low-skill region would oppose patient mobility (unless there is an interregional income transfer system in place).

The *policy implications* that can be drawn from our analysis are two-fold. First, patient mobility is beneficial for global (interregional) welfare when regions differ in their health care quality. However, patient mobility might reduce regional welfare because of the 'race-to-the-bottom' effect in terms of health care quality or the high income taxation that is needed to fund high-quality treatment to patients travelling across regions. In a decentralised regime with separate jurisdictions, the scope for patient mobility is reduced to the locus of situations that result in (weak) Pareto improvements for the regions.

Second, the success of imposing patient mobility (like the new EU legislation) crucially depends on the transfer payment system. In absence of any payments, the outcome is either no mobility, where patients seeking cross-border care are denied access, or mobility with a 'race-to-the-bottom' effect on health care quality. In presence of a payment system, the transfer price needs to be sufficiently high, so that the provider is compensated from the extra cost of treating patients from another region. Otherwise, the provider would refuse to treat patients or reduce quality in order to lower demand from mobile patients. However, the price cannot be too high either, as this would imply high tax

rates to finance the high-quality treatment to patients demanding cross-border care.

The new EU legislation states that patients seeking care in another member state are entitled to reimbursement covered by the health insurance plan in the patients' home country. However, the design of the transfer payment system is not specified by the new EU law, but to a large extent left to the member states to decide. Based on our study, we expect to see patient mobility occurring only between countries with a (weakly) mutual benefit from this trade of health care services. If the EU is enforcing the rights for patients to obtain health care in another country, without establishing a proper transfer payment system, the impact on health care quality and financing might be detrimental not just to regional welfare but also to global (interregional) welfare.

By way of conclusion, we would like to point out some limitations of our study. First, we focus on patient mobility motivated by differences in the health care quality across regions. While we think this is a main motivation for (planned) cross-border health care demand, there might be other sources as well. For instance, patients might travel to another country for treatments that are not available in their home country. Notably, the new EU law says that the patients are not eligible to reimbursement for treatments not covered by the health insurance plan in the home country, implying that such treatments would have to be covered out-of-pocket. This kind of cross-border health care demand has been in place for many years, but seems to be of limited scope according to the very low figures of patient mobility. We have therefore not addressed this issue in our paper.

Second, we have assumed that the income level is the same across and within regions, and that utility is linear in income (i.e., patients are risk neutral). As long as utility is linear in income, allowing for regional differences in income would not play any role for our analysis. The reason is that patient mobility, as defined by the marginal patient, does not depend on differences in income levels. Notably, we could re-interpret the fraction of mobile (immobile) patients as the rich (poor) patients, which are able (not able) to cover the cost of seeking care in another region. Mobility would then be higher (lower), the higher (lower) the share of rich patient living in the low-skill region. However, we find this to be a rather trivial result. Introducing income differences into the model should be done in a more elaborate way, especially by allowing for utility to be concave in income (i.e., risk-averse patients). We could also allow for different income distributions in the two regions. While this could be an interesting study, it is well beyond the scope of the current paper and thus left for future research.

References

[1] Aiura, H., Sanjo, Y., 2010. Privatization of local public hospitals: effect on budget, medical service quality, and social welfare. International Journal of Health Care Finance & Economics, 10, 275–299.

[2] Barros, P.P., Martinez-Giralt, X., 2002. Public and private provision of health care. Journal of Economics & Management Strategy, 11, 109–133.

[3] Brekke, K.R., Siciliani, L., Straume, O.R., 2011. Hospital competition and quality with regulated prices. Scandinavian Journal of Economics, 113, 444–469.

[4] Commission of the European Communities, 2006. Communication from the commission. Consultation regarding community action on health services. Brussels, September 26, 2006, SEC(2006) 1195/4.

[5] Cremer, H., Gahvari, F., 2000. Tax evasion, fiscal competition and economic integration. European Economic Review, 44, 1633–1657.

[6] Gaynor, M., 2006. What do we know about competition and quality in health care markets? Foundations and Trends in Microeconomics 2 (6).

[7] Gravelle, H., Sivey, P., 2010. Imperfect information in a quality-competitive hospital market. Journal of Health Economics, 29, 524–535.

[8] Kanbur, R., Keen, M., 1993. Jeux sans frontières: tax competition and tax coordination when countries differ in size. American Economic Review, 83, 877–892.

[9] Lyon, T.P., 1999. Quality competition, insurance, and consumer choice in health care markets. Journal of Economics & Management Strategy, 8, 545–580.

[10] Nielsen, S.B., 2001. A simple model of commodity taxation and cross-border shopping. Scandinavian Journal of Economics, 103, 599–624.

[11] Oates, W.E., 1999. An essay on fiscal federalism. Journal of Economic Literature, 37, 1120–1149.

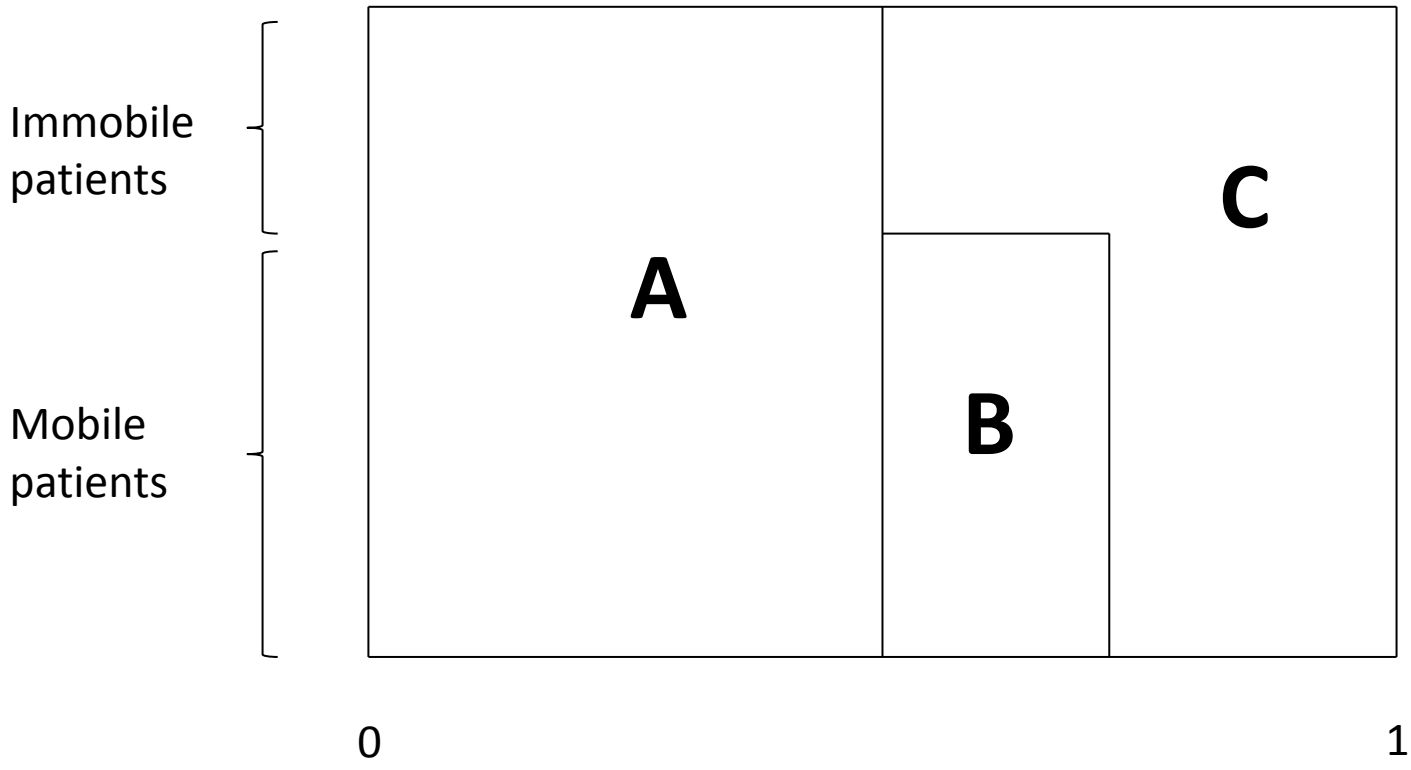
[12] Oates, W.E., 2001. Fiscal competition and European Union: contrasting perspectives. Regional Science and Urban Economics, 31, 133–145.

[13] Petretto, A., 2000. On the cost-benefit of the regionalisation of the National Health Service. Economics of Governance, 1, 213–232.

[14] Trandel, G.A., 1994. Interstate commodity tax differentials and the distribution of residents. Journal of Public Economics, 53, 435–457.

[15] Wang, Y.-Q., 1999. Commodity taxes under fiscal competition: Stackelberg equilibrium and optimality. American Economic Review, 89, 974–981.

Figure 1. Effect of patients' mobility when transfer payment is equal to marginal cost



Patients in area A are indifferent
Patients in area B gain
Patients in area C lose