Public provision and cross-border health care[1]

David Granlund[1,2] and Magnus Wikström[1]

[1]Department of Economics, Umeå University, SE-901 87 Umeå, Sweden
[2]HUI Research, SE-103 29 Stockholm, Sweden.

October 9, 2013

Abstract:
We study how the optimal public provision of health care depends on whether or not individuals have an option to seek publicly financed treatment in other regions. We find that, relative to the first-best solution, the government has an incentive to over-provide health care to low-income individuals. When cross-border health care takes place, this incentive is solely explained by that over-provision facilitates redistribution. The reason why more health care facilitates redistribution is that high-ability individuals mimicking low-ability individuals benefit the least from health care when health and labor supply are complements. Without cross-border health care, higher demand for health care among high-income individuals also contributes to the over-provision given that high-income individuals do not work considerably less than low-income individuals and that the government cannot discriminate between the income groups by giving them different access to health care.

Keywords: Health expenditure; Income redistribution; Patient mobility; Public Provision; Waiting time.
JEL classification: H42; H51; I11; I18

1The authors would like to thank Thomas Aronsson for helpful comments and suggestions. We also acknowledge research grants from the Swedish Council for Working Life and Social Research and jointly from Bank of Sweden Tercentenary Foundation, the Swedish Council for Working Life and Social Research, and the Swedish Tax Agency.

1. Introduction

In a publicly financed health care system, cross-border health care means that individuals, under certain circumstances, can get access to publicly financed health care outside of their region or country of residence. There are several reasons why cross-border care may be beneficial. Since some health care services are characterized by increasing returns to scale, specialization of certain treatments may lower cost and increase quality. This is especially true for activities that require large capital investments and special competence. Another important reason is that waiting times vary between regions, and cross-border care can help to bring down waiting times in regions with relatively high demand and low capacity. This paper discusses the optimal provision of public health care when cross-border care is driven by differences in waiting times between regions.

Long waiting times are predominantly found in health care systems with tax-financed insurance and a global expenditure budget, as compared to fee for service payments (Cullis, Jones and Propper, 2000). Many queues are constantly too long for them to be explained only as a tool to increase capacity utilization when the demand for health services fluctuates over time (Hoel and Saether, 2003). Instead, waiting times are often seen as a rationing device (Cullis, Jones and Propper, 2000). Lindsay and Feigenbaum (1984) suggest that people might incur cost for joining a queue. When these immediate costs are compared to the discounted benefits of treatments, long waiting times can induce some patients never to join the queue. Also, physicians might be less inclined to list patients as waiting for treatment if the queues are long enough (Cullis, Jones and Propper, 2000).[2] The number of people exiting queues before treatment can be affected by longer queues leading to more people recovering before treatment or adjusting to health defects so that they no longer find it unpleasant enough to warrant treatment (Hoel and Saether, 2003). In the extreme case, longer queues might also increase the number of exits since more people might die before treatment, which Plump et al. (1999) found for cardiac surgery in The Netherlands. Hoel and Saether (2003) emphasize that people may choose private health care instead if the public queues are too long.

Increasing budgets is one policy that can be used to reduce waiting times. The effect of a budget increase is, however, reduced if it affects the inflow and outflow before treatment. Since waiting times differ between regions, another policy option is to allow people to travel

[2] Martin and Smith (1999), however, report that the elasticity of demand for elective surgery with respect to waiting time is low.

across borders in order to obtain treatment. This is a policy currently discussed within the EU. As of March 2011, a directive (EU directive 2011/24) was addressed to the member states concerning patients' rights in cross-border health care. Although the rules regarding cross-border care still are left to be settled among the member states, and the fact that cross-border care currently comprises only about one percent of member states' total public health care budgets, the directive can be considered a step towards a well-functioning cross-border care system.[3]

In this paper we study how the optimal provision of health care depends on a policy to increase accessibility by the use of cross-border health care. We assume that individuals have a possibility to seek treatment in regions with shorter waiting times and study how this affects the optimal size of the health care budget in their home region, which in turn affects the waiting times there. In particular, we focus on the provision towards individuals who do not use cross-border care by studying how the incentives for over-provision change as cross-border care becomes available. We define over-provision as providing so much health care so that the marginal cost exceeds the marginal benefit; that is, providing more health care than the first-best solution would stipulate. One of the central goals in public health care is equal access to health care among the inhabitants, which means that it should not be possible to discriminate care between individuals based on other than medical reasons. Therefore, we take as a starting point for our analysis that the public sector cannot discriminate between individuals by having different waiting times for different individuals.

Following Stern (1982) and Stiglitz (1982), we set up a model with two types of individuals which differ in ability and thus in the wage rate they earn. As emphasized in earlier research (Blomquist and Christiansen, 1995, 1998; Boadway and Marchand, 1995), public provision of private goods, of which health care is one example, can be an important tool for redistribution. To study the redistributive role of health care provision, we follow this literature in assuming that the government cannot observe individuals' ability-types, only their gross incomes. This constrains redistribution since high-ability individuals can choose to mimic low-ability individuals by working less if the tax payments on high incomes become too high. The model contains two regions and we assume that the government in each region has access to a non-linear income tax system which facilitates redistribution between the

[3] Waiting times can also be reduced by giving providers of health care financial incentives to do so (Siciliani and Hurst, 2005). Such financial incentives can also be strengthen by policies such as waiting time guarantees or by granting patients the right to freely choose hospital.

ability types. One of the regions contains a larger fraction of the high-ability type resulting in higher health expenditures and shorter waiting times. High-ability types in the region with a relatively large fraction of low-ability types may then want to utilize cross-border care since it may reduce their waiting time.

The paper contributes to the existing literature regarding waiting times by analyzing the government's incentive for health care provision and waiting times when it has access to efficient means of redistribution and by illustrating that reducing waiting times can facilitate further redistribution. First, the solution is characterized when cross-border care is not available. This solution generally implies over-provision of care to low-ability type individuals since the optimal policy will be an (weighted) average of low- and high-type demand and since the possibility of mimicking by high-types will keep waiting times down as long as labor supply and health are complements. Then we discuss what the optimal policy will look like when cross-border care is allowed for. In our model only citizens in one of the regions may find it beneficial to use cross-border care, while at the same time assumptions are made so that the receiving regions' policy is independent of whether cross-border care takes place or not. Within this framework then, two situations are illustrated; one in which the sending region finds it beneficial to utilize cross-border care, and one in which it doesn't. In both cases, we find that there are incentives for over-provision towards low-type individuals although for different reasons. If the government finds it beneficial to promote cross-border care, the mimicking self-selection constraint may still be binding, which induces over-provision. In the other case, the government does not want high-ability individuals to travel for care, and the binding constraint in this case is the utility difference between going and staying. Therefore, there will be an incentive to keep waiting times lower than if cross-border care was not available.

The rest of the paper is organized as follows: section 2 sets up the basic model to be used in the paper. In section 3, we discuss optimal health care provision when cross-border health care is not available, while section 4 discusses optimal provision when cross-border health care is an option. Finally, section 5 concludes.

2. The model

The federation consists of two regions which are responsible for providing health care. The regions finance their expenditure via a non-linear labor income tax and balance their budgets.

Each region is populated by two ability types; a low-ability type (denoted by superindex 1) and a high-ability type (denoted by superindex 2). Let the wage rate earned by ability type *i* be w^i. The high-ability type is more productive and therefore earns a higher wage rate, $w^2 > w^1$. The distribution of ability-types differs between regions. We denote the region with the relatively higher share of low-ability individuals by *L* and the region with the relatively higher share of high-ability individuals by *H*. The proportion of ability-type *i* in region *j* is denoted γ_j^i. To simplify the notations, we normalize the population in each region to one. We disregard migration, but residents are assumed to be able to seek health care in the other region.

The utility facing ability-type *i* living in region *j* is given by $U_j^i = U(x_j^i, z_j^i, h_j^i)$, where *x* is private consumption, *z* leisure, and *h* is time as healthy. Leisure is defined as a time endowment, normalized to one, less the time spent in market work, *l*. Preferences are identical and the utility is increasing and strictly concave in all arguments. We also assume that health and labor supply are complements in the utility function. This assumption is equivalent to assuming that sickness increases the marginal disutility of work.

Health care is assumed to be supplied free of charge. We also assume that the regions are not allowed to discriminate between low and high-ability types and between residents and non-residents when providing health care. Thus, all individuals seeking treatment in a region will have the same waiting time. The share of time as sick for ability-type *i* living in region *j* that seeks treatment in region *k* is written

$$a_j^i = \frac{N_k - e_k}{rN_k} \tag{1}$$

where N_k is the number that seeks treatment in region *k*.[4] The variable e_k denotes the number of individuals who get medical treatment; all individuals who get treatment are assumed to be recovered. Lastly, *r* denotes the exogenous rate at which individuals recover by themselves. Note that the normalization that all individuals fall sick imply that r>1 since a_j^i otherwise would be equal or larger than unity without treatments. From equation (1) we see that $\partial a_j^i / \partial e_k = -1/rN_k < 0$.

4 Equation (1) is the steady state solution to the dynamic equation $\dot{a}_j^i = N_k - N_k r a_j^i - e_k$, where $\dot{a}_j^i$ is the change in a_j^i.

To simplify the notation, we assume that all individuals fall sick one time, meaning that N_k=1 without cross-border health care. With this assumption, a_j^i also equals the average sickness spell for ability-type *i* from region *j*, and we also assume that a_j^i equals the actual sickness spell for all individuals of ability-type *i* from region *j*. Thus, we neglect intra-type intra-regional heterogeneity, that otherwise would occur since some individuals recover by themselves, to better be able to analyze the heterogeneity we are interesting in – that between regions and ability-types. The assumption that a_j^i equals the sickness spell for all individuals of ability-type *i* from region *j* implies that $h_j^i = 1 - a_j^i$. We also simplify the analysis by assuming that the cost of seeking treatment in the other region is realized for all individuals who have joined the queue in the other region, including those who recover by themselves.

We assume that both regions have the same cost function for treatments, *c(e)* and that the marginal cost is positive and increasing (i.e. $c'>0$ and $c''>0$). The governments' total cost for health care will be lower the longer the waiting time is; with a positive waiting time, some will recover before it is their time to get treatment, meaning that not all who fall sick have to be treated. With a waiting time of Q_k^i, $N_k r^{Q_k^i}$ will recover before getting treatment. Like a_j^i, Q_k^i is a decreasing function of e_k.[5]

The private agents are assumed to make their choices concerning labor supply and where to seek treatment after the governments' policies have been proclaimed. The private agents maximize their utility, subject to their time constraints, $1-l_j^i = z_j^i$, and their budget constraints, $w^i l_j^i - T_j(w^i l_j^i) = x_j^i + b\delta_j^i$, where *T* is the income tax payment. The parameter *b* describes the (travel) cost of seeking treatment in another region than that of residence and δ_j^i is a dummy that takes the value one if the individual actually seeks treatment in another region. By defining the marginal net wage as $\widetilde{w}_j^i = w^i\left[1 - \partial T_j/\partial\left(w^i l_j^i\right)\right]$, the first order condition for consumption and labor supply can be written $\frac{\partial U_j^i}{\partial x_j^i}\widetilde{w}_j^i = \frac{\partial U_j^i}{\partial z_j^i}$.

[5] Note that $N_k r^{Q_k^i} = N_k r a_j^i$, since both expressions define the numbers that recover themselves, which gives $Q_k^i \ln r = \ln r + \ln a_j^i$ and $Q_k^i = 1 + \ln a_j^i / \ln r$. Hence, $\partial Q_k^i/\partial a_j^i = \ln r/a_j^i > 0$ (since r>1) and $\partial Q_k^i/\partial e_k = -\ln r/\left(r N_k a_j^i\right) < 0$.

3. Government policy without cross-border health care

We assume that the governments have Utilitarian social welfare functions that can be written

$$W_j = \omega_j^1 \gamma_j^1 U_j^1\left(x_j^1, z_j^1, h_j\right) + \omega_j^2 \gamma_j^2 U_j^2\left(x_j^2, z_j^2, h_j\right) \quad \mathrm{j} = L, H, \tag{2}$$

where ω_j^i denotes the weights government *j* attaches to the welfare of each ability-type. Note that, in the absence of cross-border health care, both ability types face the same waiting time, meaning that $h_j^1 = h_j^2 = h_j$. The government budget constraint in region *j* is given by

$$\sum_i \gamma_j^i \left(w^i l_j^i - x_j^i\right) = c(e_j). \tag{3}$$

We make the conventional assumptions about information; governments can observe income, whereas ability is private information. We follow much of the earlier literature in concentrating on a standard case, where the governments want to redistribute from the high-ability to the low-ability type. As a consequence, they would like to prevent the high-ability type from pretending to be a low-ability type, i.e. becoming a mimicker. This is accomplished by imposing a self-selection constraint, implying that the high-ability type (at least weakly) prefers the combination of disposable income and hours of work intended for him/her over the combination intended for the low-ability type. Note that the hours of leisure that the high-ability type enjoys when working just enough to reach the same labor income as the low-ability type is given by $\hat{z}_j^2 = 1 - l_j^1(w^1/w^2)$ and that the self-selection constraint can be written

$$U_j^2\left(x_j^2, z_j^2, h_j\right) \geq \hat{U}_j^2\left(x_j^1, \hat{z}_j^2, h_j\right). \tag{4}$$

Each regional government decides on a non-linear income tax schedule, $T_j(w^i l_j^i)$, and the number of treatments, e_j. The choice of e_j in turn affects the sickness spell a_j: without cross-border health care $a_j = \left(1 - e_j\right)/r$. Since the government has access to a non-linear tax-schedule, we can solve the policy problem as if it directly chooses x_j^1, l_j^1, x_j^2, l_j^2 and e_j to maximize its objective function, subject to the self-selection constraint (4) and the resource constraint. The Lagrangean corresponding to the optimization problem facing the government is written

$$\begin{aligned}\pounds_j = {} & \gamma_j^1 \omega_j^1 U_j^1\left(x_j^1, z_j^1, h_j\right) + \gamma_j^2 \omega_j^2 U_j^2\left(x_j^2, z_j^2, h_j\right) \\ & + \lambda_j\left[U_j^2\left(x_j^2, z_j^2, h_j\right) - \widehat{U}_j^2\left(x_j^1, \hat{z}_j^2, h_j\right)\right] \\ & + \mu_j\left[\gamma_j^1 w^1 l_j^1 + \gamma_j^2 w^2 l_j^2 - \gamma_j^1 x_j^1 - \gamma_j^2 x_j^2 - c\left(e_j\right)\right]\end{aligned} \tag{5}$$

The first order conditions become

$$\frac{\partial \pounds_j}{\partial x_j^1} = \gamma_j^1\left[\omega_j^1 \frac{\partial U_j^1}{\partial x_j^1} - \mu_j\right] - \lambda_j \frac{\partial \widehat{U}_j^2}{\partial x_j^1} = 0 \tag{6}$$

$$\frac{\partial \pounds_j}{\partial x_j^2} = \gamma_j^2\left[\omega_j^2 \frac{\partial U_j^2}{\partial x_j^2} - \mu_j\right] + \lambda_j \frac{\partial U_j^2}{\partial x_j^2} = 0 \tag{7}$$

$$\frac{\partial \pounds_j}{\partial l_j^1} = -\gamma_j^1\left[\omega_j^1 \frac{\partial U_j^1}{\partial z_j^1} - \mu_j w^1\right] + \lambda_j \frac{w^1}{w^2} \frac{\partial \widehat{U}_j^2}{\partial \hat{z}_j^2} = 0 \tag{8}$$

$$\frac{\partial \pounds_j}{\partial l_j^2} = -\gamma_j{}^2\left[\omega_j{}^2 \frac{\partial U_j{}^2}{\partial z_j{}^2} - \mu_j w^2\right] - \lambda_j \frac{\partial U_j{}^2}{\partial z_j{}^2} = 0 \tag{9}$$

$$\frac{\partial \pounds_j}{\partial e_j} = \left[\gamma_j^1 \omega_j^1 \frac{\partial U_j^1}{\partial h_j} + \gamma_j^2 \omega_j^2 \frac{\partial U_j^2}{\partial h_j} + \lambda_j\left(\frac{\partial U_j^2}{\partial h_j} - \frac{\partial \widehat{U}_j^2}{\partial \hat{z}_j^2}\right)\right]\frac{1}{r} - \mu_j c' = 0 \tag{10}$$

Note first, that the first-order conditions for consumption and labor supply (6-9) are standard in the optimal taxation literature. This also implies that the optimal taxation formulas that can be derived follow the standard pattern; see e.g. Aronsson and Blomquist (2008). The term *1/r* in the first order condition (10) is the negative of $\partial a_j/\partial e_j$ and thus describes how the sickness spell is decreased when the number of treatments is increased marginally. Put differently, *r* is the number that treatments must be increase by in order to reduce the sickness spell with one unit. The smaller *r* is, the higher the optimal level of e_j will be.

Defining $MRS_j^i = \frac{\partial U_j^i/\partial h_j^i}{\partial U_j^i/\partial x_j^i}$ to be the marginal rate of substitution between health and private consumption and combining first order conditions (6), (7) and (10) gives us the following proposition:

Proposition 1: *Without cross-border health care, the rule for optimal provision of health care is*

$$rc' = \sum_i \gamma_j^i\, MRS_j^i + \frac{\lambda_j}{\mu_j}\frac{\partial \widehat{U}_j^2}{\partial x_j^1}\left[MRS_j^1 - \widehat{MRS_j^2}\right]. \tag{11}$$

The proposition states that the number of treatments, e_j, should be chosen such that the marginal cost of additional time as healthy, rc', equals a weighted average of the marginal rates of substitution between health and consumption plus a term associated with the self-selection constraint. As mentioned above, health and labor supply are assumed to be complements, meaning that the term in square brackets is positive since the mimicker supplies less labor than the true low-ability type. This means that increasing the number of treatments slightly (or equivalently reducing the waiting time) relaxes the self-selection constraint. Thus, the proposition implies that reducing the waiting time will allow the governments to redistribute more from the high-ability type to the low-ability type.

Hoel and Saether (2003) derived a different result, stating that longer waiting times can serve as a means of redistribution. The intuition behind their result was that patients with low waiting cost are better off waiting, since waiting time makes patients with high waiting costs choose private treatment which reduces the cost of publicly financed health care. The difference in results is explained by that we take into account how mimickers, and hence the self-selection constraint, are affected by waiting times and that mimickers are those that have the lowest waiting costs.[6] The latter follows from the assumption that health and labor supply are complements and the result would hold even if private health care without waiting times were introduced in this model.[7] It should, however, be made clear that waiting times can serve as a means of redistribution in the way explained by Hoel and Saether if the government knows that the self-selection constraint is not binding and prefers to redistribute more but is unable to do so using efficient means such as non-linear income taxation; for example, if the government is restricted to using a proportional income tax or a head tax.[8]

Using that $\gamma_L^1 - 1 = -\gamma_L^2$, equation (11) can be rewritten as

[6] Fossati and Levaggi (2008) also derived a result different from that of Hoel and Saether (2003). They used a model where public health care was financed through a linear income tax and where private treatment could be bought without delay.

[7] With private health care bought by the true high-ability type, so that publicly provided health care is only directed towards the low-income consumers, equation (11) would change to $rc' = MRS_j^1 + \frac{\lambda_j}{\gamma_j^1 \mu_j} \frac{\partial \hat{U}_j^2}{\partial x_j^1} \left[MRS_j^1 - \widehat{MRS_j^2} \right]$, meaning that the result that shorter waiting time enables more redistribution would remain unchanged.

[8] Besley and Coate (1991) analyze situations under which a low quality of a publicly provided private good might facilitate redistribution.

$$rc' - MRS_j^1 = \gamma_j^2\left(MRS_j^2 - MRS_j^1\right) + \frac{\lambda_j}{\mu_j}\frac{\partial \widehat{U}_j^2}{\partial x_j^1}\left[MRS_j^1 - \widehat{MRS}_j^2\right]. \quad (12)$$

Note that rc', the marginal cost of additional time as healthy, equals the marginal rate of transformation between health and private consumption, since the price of private consumption is normalized to unity. Thus, the left hand side of equation (12) shows for the low-ability type the difference between the marginal rate of transformation and the marginal rate of substitution, a difference that in a first best scenario would be zero.

As discussed above, the term in square brackets is positive, thus showing one incentive to over-provide health care to the low-ability type, i.e. to have shorter waiting times for them than would be optimal in a first best scenario. Assuming that $l_j^2 \geq l_j^1$, the first term on the right hand side of equation (12) is also positive,[9] indicating a second reason for over-provision of health care to the low-ability type. That is, the waiting time for the low-ability type will be held down by the fact that the first best waiting time for the high-ability type is lower than the first best waiting time for the low-ability type.

In region L where γ_j^2 is lower than in region H, the first term on the right hand side of equation (12) will be smaller than in region H, i.e., health care will be less over-provided to the low-ability type in region L. This is one important reason to why, without cross border health care, e_j is likely decreasing in γ_j^1 meaning that e_L likely is lower than e_H. As discussed in the Appendix, the sign of $de_j/d\gamma_j^1$ is also affected by indirect effects and is generally not possible to determine, but in the following we assume that $de_j/d\gamma_j^1 < 0$.

4. The government's problem with cross-border health care

In this section we discuss the behavior of governments when individuals can choose the region of treatment. We assume that the home region pays the incremental treatment cost to the other region when someone seeks treatment in that region, which implies that only the incentives for the region with the longest waiting time will be affected by the possibility of cross border health care. First, we discuss the case when the high-ability type from region L

[9] Note that, without cross border health care, both types have the same health while the high ability-type has a larger private consumption. Thus, unless l_j^1 is much larger than l_j^2, which due to the complementarities between labor supply and health could result in the low-ability type having higher marginal utility of health, this implies that $MRS_j^1 < MRS_j^2$. Thus, $l_j^2 \geq l_j^1$ is a sufficient but not a necessary condition for $MRS_j^1 < MRS_j^2$.

actually seeks treatment in region H. That the low-ability workers will never be the worker type choosing treatment in region H follows by the assumption that $l_j^2 \geq l_j^1$, which guarantees that $MRS_j^1 < MRS_j^2$. Then, we discuss the case when government L induces its citizens to not to seek treatment outside their own jurisdiction. The latter situation might arise for example when a federal government or a supranational organization like the European Union have given individuals the right to cross border health care even if it is not in the interest of all regions.

When cross-border health care takes place

Cross border health care takes place when it is welfare improving for region L, which it can be since it provides the government in region L with the possibility to let the high-ability type wait a shorter time for treatment than the low-ability type. Different waiting times are beneficial for the government, since the cost of treatment is the same for both types while the marginal rate of substitution between health and consumption, and thus the waiting cost, is larger for the high-ability type. Since the government can redistribute using non-linear income taxes, it might in fact be welfare enhancing for both ability types. However, since the government in region L cannot choose the waiting time in region H, and because of travel costs and increasing marginal cost of treatments in the two regions, cross border health care can never achieve as high utility as if the governments were allowed to have different waiting times for the two ability types. Thus, an underlying assumption made here is that it is not feasible for the government to implement a policy in which there are two separate health care ques.

Note that if cross border health care is welfare enhancing, it is always possible to make the high-ability type seek treatment in the other region. The reason is that they will only pay part of the cost for it, which is the travel cost but not the increased cost of treatment, but they will get the entire benefit from it conditional on x_L^2 and l_L^2.

Let us now analyze the effect of cross border health care on government policy. The high-ability type may choose to mimic the low-ability type or choose treatment in the other region. However, the high-ability type would not choose treatment in the other region and at the same time mimic the low-ability type, since she would then have equal income but lower marginal utility of health than the low-ability type. The Lagrange function for the policy problem in

region L is given by

$$\begin{aligned}\pounds &= \gamma_L^1\omega_L^1 U_L^1\left(x_L^1, z_L^1, h_L^1\right)+\gamma_L^2\omega_L^2 U_L^2\left(x_{L,b}^2, z_L^2, h_{L,b}^2\right)\\ &+\lambda_L\left[U_{L,b}^2\left(x_{L,b}^2, z_L^2, h_{L,b}^2\right)-\hat{U}_L^2\left(x_L^1, \hat{z}_L^2, h_L^{2}\right)\right]\\ &+\mu_L\left[\gamma_L^1 w^1 l_L^1+\gamma_L^2 w^2 l_L^2-\gamma_L^1 x_L^1-\gamma_L^2\left(x_{L,b}^2+b\right)-p(e_L)-P(e_H)\right],\end{aligned} \tag{13}$$

where $x_{L,b}^2 = x_L^2 - b$ and $h_{L,b}^2 = 1 - a_H$ describe the outcomes for high-ability types seeking treatment abroad and where $P(e_H)$ is the total payment to region H for residents of region L that are treated in H.

Note that $h_L^1 = \hat{h}_L^2 = 1 - a_L = 1 - (\gamma_L^1 - e_L)/\gamma_L^1 r$ depends on e_L, while $h_L^2 = 1 - a_H = 1 - \left(1 + \gamma_L^2 - e_H\right)/\left[\left(1+\gamma_L^2\right)r\right]$ is unaffected by marginal changes in region L's decisions. The first order conditions can be written

$$\frac{\partial \pounds_L}{\partial x_L^1} = \gamma_L^1\left[\omega_L^1\frac{\partial U_L^1}{\partial x_L^1} - \mu_L\right] - \lambda_L\frac{\partial \hat{U}_L^2}{\partial x_L^1} = 0 \tag{14}$$

$$\frac{\partial \pounds_L}{\partial x_L^2} = \gamma_L^2\left[\omega_L^2\frac{\partial U_{L,b}^2}{\partial x_{L,b}^2} - \mu_L\right] + \lambda_L\frac{\partial U_{L,b}^2}{\partial x_{L,b}^2} = 0 \tag{15}$$

$$\frac{\partial \pounds_L}{\partial l_L^1} = -\gamma_L^1\left[\omega_L^1\frac{\partial U_L^1}{\partial z_L^1} - \mu_L w^1\right] + \lambda_L\frac{w^1}{w^2}\frac{\partial \hat{U}_L^2}{\partial \hat{z}_L^2} = 0 \tag{16}$$

$$\frac{\partial \pounds_L}{\partial l_L^2} = -\gamma_L^2\left[\omega_L^2\frac{\partial U_{L,b}^2}{\partial z_L^2} - \mu_L w^2\right] - \lambda_L\frac{\partial U_{L,b}^2}{\partial z_L^2} = 0 \tag{17}$$

$$\frac{\partial \pounds_L}{\partial e_L} = \left[\gamma_L^1\omega_L^1\frac{\partial U_L^1}{\partial h_L^1} - \lambda_L\frac{\partial \hat{U}_L^2}{\partial h_L^1}\right]\frac{1}{\gamma_L^1 r} - \mu_L c' = 0 \tag{18}$$

If these first-order conditions are compared with those obtained when cross-border health care was not allowed, (6)-(10), we note that the first order conditions for x_L^1 and l_L^1 are unchanged, while the other three conditions are different. The high-ability type's payment for travel cost increases his/her marginal benefit of consumption, which has a positive effect on the choice of x_L^2. His/her marginal utility of leisure is reduced (since health and labor supply are assumed to be complements) which has a positive effect on the choice of l_L^2. The largest difference is that the high-ability type no longer is directly affected by marginal changes in the choice of

e_L. To see how the incentives for health care provision are changed by cross-border health care, the first order conditions (14) and (18) are combined. The following result applies

Proposition 2: *When cross-border health care takes place, the rule for optimal provision of health care in region L is written*

$$rc' = MRS_L^1 + \frac{\lambda_L}{\gamma_L^1 \mu_L} \frac{\partial \widehat{U}_j^2}{\partial x_j^1} \left[MRS_L^1 - \widehat{MRS_L^2} \right]. \tag{19}$$

First note that Proposition 2 implies that the government in region L has an incentive to over-provide health care to the low-ability type also in the case when only the low-ability type is treated within the region. The reason is that the high-ability type would seek treatment in region L if she chooses to mimic the low-ability type and value short waiting time less than the true low-ability type.

Over-provision to the low-ability type is reduced by the fact that MRS_j^2 does not affect the rule for optimal provision in this case. On the other hand, the importance of the self-selection term is scaled up with the factor $1/\gamma_L^1$, compared with equation (11). The intuition is that over-provision to relax the self-selection constraint no longer affects the entire population (normalized to unity), only the share γ_L^1 of the population that seeks treatment in the home region.

The actual number of treatments provided to the low-ability does not only depend on the degree of over-provision. For a given degree of over-provision, the number of treatments to the low-ability type is larger with cross border health care since the marginal cost of treatments in region L is smaller when the high-ability type no longer requires health care resources from that region. However, cross-border health care also means that the cost of health care provision for the high-ability type implies an income effect, which contributes towards a decrease in the number of treatments in the home region. As discussed in the beginning of this section, if cross border health care is welfare improving, it is likely to improve welfare for both ability-types since the government has the possibility to redistribute resources using income taxes. This holds irrespective of whether the waiting time for the low-ability type is reduced or increased.

The government prevents cross-border health care

If the travel cost, *b*, and the second derivative of the cost function, *c´´*, are sufficiently large, and the waiting time in region H is too far away from what region L would prefer for its high-ability type, then cross-border health care will reduce welfare in region L. For example, if the ability distribution is very similar across the regions, then the waiting times will also be very similar, implying that cross-border health care will not be welfare improving even if *b* and *c´´* are quite low. If the welfare losses associated with cross-border health care are sufficiently large relative to the distortions necessary to prevent it, the government in region L will induce the high-ability type not to seek treatment abroad.

To prevent the high-ability type from seeking treatment in region H, the government in region L makes sure that the following self-selection constraint is fulfilled

$$U_L^2\left(x_L^2, z_L^2, h_L^2\right) \geq U_{L,b}^2\left(x_{L,b}^2, z_L^2, h_{L,b}^2\right), \tag{21}$$

The Lagrange becomes

$$\begin{aligned} £ &= \gamma_L^1 \omega_L^1 U_L^1\left(x_L^1, z_L^1, h_L^1\right) + \gamma_L^2 \omega_L^2 U_L^2\left(x_L^2, z_L^2, h_L^2\right) \\ &+ \lambda_L\left[U_L^2\left(x_L^2, z_L^2, h_L^2\right) - \hat{U}_L^2\left(x_L^1, \hat{z}_L^2, h_L^2\right)\right] \\ &+ \eta_L\left[U_L^2\left(x_L^2, z_L^2, h_L^2\right) - U_{L,b}^2\left(x_{L,b}^2, z_L^2, h_{L,b}^2\right)\right] \\ &+ \mu_L\left[\gamma_L^1 w^1 l_L^1 + \gamma_L^2 w^2 l_L^2 - \gamma_L^1 x_L^1 - \gamma_L^2 x_L^2 - p(e_L)\right]. \end{aligned} \tag{22}$$

Note that only one of the self-selection constraints can be binding, since $U_{L,b}^2\left(x_{L,b}^2, z_L^2, h_{L,b}^2\right)$ cannot be both strictly below and strictly above $\hat{U}_L^2\left(x_L^1, \hat{z}_L^2, h_L^2\right)$. If the first self-selection constraint is binding, the outcome will be identical with that discussed above when cross-border health care was not allowed. Therefore, we concentrate below on the case when the "no cross-border health care" constraint, (21), is binding. In this case, the first order conditions become

$$\frac{\partial £_L}{\partial x_L^1} = \gamma_L^1\left[\omega_L^1 \frac{\partial U_L^1}{\partial x_L^1} - \mu_L\right] = 0 \tag{23}$$

$$\frac{\partial £_L}{\partial x_L^2} = \gamma_L^2\left[\omega_L^2 \frac{\partial U_L^2}{\partial x_L^2} - \mu_L\right] + \eta_L\left[\frac{\partial U_L^2}{\partial x_L^2} - \frac{\partial U_{L,b}^2}{\partial x_{L,b}^2}\right] = 0 \tag{24}$$

$$\frac{\partial £_L}{\partial l_L^1} = -\gamma_L^1\left[\omega_L^1 \frac{\partial U_L^1}{\partial z_L^1} - \mu_L w^1\right] = 0 \tag{25}$$

$$\frac{\partial £_L}{\partial l_L^2} = -\gamma_L^2 \left[\omega_L^2 \frac{\partial U_L^2}{\partial z_L^2} - \mu_L w^2 \right] - \eta_L \left[\frac{\partial U_L^2}{\partial z_L^2} - \frac{\partial U_{L,b}^2}{\partial z_L^2} \right] = 0 \tag{26}$$

$$\frac{\partial £_L}{\partial e_L} = \left[\gamma_L^1 \omega_L^1 \frac{\partial U_L^1}{\partial h_L} + \left(\gamma_L^2 \omega_L^2 + \eta_L \right) \frac{\partial U_L^2}{\partial h_L} \right] \frac{1}{r} - \mu_L c' = 0 \tag{27}$$

Comparing these first order conditions with those that pertained when cross-border health care was not allowed, (6)-(10), we see some distinct differences. Looking at the new terms, which are present in the first order conditions that directly affects the high-ability type, (24), (26) and (27), they work in the direction of a lower x_L^2, since $\left[\frac{\partial U_L^2}{\partial x_L^2} - \frac{\partial U_{L,b}^2}{\partial x_{L,b}^2} \right] < 0$, a lower l_L^2, since $\left[\frac{\partial U_L^2}{\partial z_L^2} - \frac{\partial U_{L,b}^2}{\partial z_L^2} \right] > 0$, and a higher e_L, since $\frac{\partial U_L^2}{\partial h_L} > 0$.[10] The intuition for a lower x_L^2 is that the government can make cross-border health care less appealing by lowering the net income for the high-ability type, since the high-ability type will have a more tight budget if he/she seeks treatment abroad, due to having to pay the travel cost. In other words, with lower income, the high-ability type will find cross border treatment to be worth less in terms of foregone consumption. Similarly, a lower l_L^2 and a higher e_L reduce the high-ability type's marginal utility of health, since the marginal utility of health is decreasing in both leisure and health itself. A higher e_L also affects the decision by reducing the difference in waiting time between the regions and hence the amount of additional health that could be achieved by cross-border health care.

Combining first order conditions (23), (24) and (27) gives us the proposition:

Proposition 3: *When cross-border health care is allowed but the government in region L prevents it, the rule for optimal provision of health care in region L is written*

$$rc' = \sum_i \gamma_J^i MRS_L^i + \frac{\eta_L}{\mu_J} \frac{\partial U_{L,b}^2}{\partial x_{L,b}^2} MRS_L^2 \tag{28}$$

[10] Those who seek care outside of their home region have to pay their travel costs, which increases their marginal utility of consumption. The first inequality holds given that this effect is not dominated by health and consumption being strong substitutes. Empirical evidence suggests that health and consumption are complements; Viscusi and Evans (1990) estimate the marginal utility of consumption when ill to be 77 percent of that when well. The corresponding estimates reported in Gilleskie (1998) for acute illness are 58 and 16, depending on the type of illness.

Proposition 3 shows that health care to the low-ability type will be over-provided also in this case, since $MRS_j^2 > MRS_j^1$ and since shorter waiting times relax the "no cross-border health care" constraint. Like proposition 1, proposition 3 states that the number of treatments, e_L, should be chosen such that the marginal cost of additional time as healthy, rc', equals a weighted average of the marginal rates of substitution between health and consumption plus a term associated with the self-selection constraint. However, equation (28) and equation (11) (in proposition 1) differ by referring to different self-selection constraints. Without further assumptions it is not possible to show if the second term of equation (28) is larger or smaller than the second term of equation (11) since the terms refers to different solutions, implying that the values of the marginal utilities and shadow prices differ. There are, however, some arguments suggesting that $\frac{\eta_L}{\mu_J}\frac{\partial U_{L,b}^2}{\partial x_{L,b}^2}MRS_L^2$ in equation (28) is larger than $\frac{\lambda_j}{\mu_j}\frac{\partial \widehat{U}_j^2}{\partial x_j^1}\left[MRS_j^1 - \widehat{MRS}_j^2\right]$ in equation (11). First, even if MRS_L^2 in equation (28) and $\left[MRS_j^1 - \widehat{MRS}_j^2\right]$ in equation (11) refer to different situations, it is quite reasonable that the first is larger than the second, since MRS_L^2 under reasonable assumptions is larger than MRS_L^1 and hence larger than MRS_L^1 minus a positive number. Also, it is likely that $\eta_L > \lambda_J$ since the utility the high-ability type could get by using cross-border care is larger than that obtained by mimicking the low-ability type.

The fact that the high-ability type now must be guaranteed a higher utility level in order to behave as the government prefers than when cross-border health care is not allowed limits the possibilities for redistribution in region L. It also reduces the total weighted welfare in the region since it implies that the government is more constrained. Thus, compared to when cross-border health care is not allowed, a low-ability worker under this situation is likely to get a smaller piece of a smaller pie and is, therefore, most likely worse off.

5. Summary and Discussion

This paper addresses optimal provision of public health care and how this is affected by the possibility of residents seeking treatment in other regions. The analysis is based on a two-type model where the regional governments use nonlinear income taxes to finance their health care cost but are constrained from applying different waiting times for the two ability types.

Consumers recover by themselves at an exogenous rate implying that the optimal waiting times are positive. However, since the ability-type distribution differs across regions, so does the waiting times, which is the driving force between cross-border health care in the model. The analysis focuses on the incentives for the region for which the optimal waiting time is longest.

The results show that it is optimal to over-provide health care to the low-ability type both when cross-border health care is allowed for in the model and when it is not allowed for. This means that there is an incentive to have shorter waiting times relative to the first-best solution. When cross border health care is not allowed for in the model, over-provision arises since this increases the possibility to redistribute without causing high-ability individuals to reduce their labor supply in order to mimic low-ability individuals and thus get their lower tax payments. In other words, health care is over-provided to relax a self-selection constraint. This result is explained by that the complementarity between health and labor supply implies that mimickers, due to their lower labor supply, value health less at the margin than low-ability individuals. Another reason for the over-provision to low-ability individuals is that the government, by assumption, is constrained from applying different waiting times in combination with that the marginal rate of substitution between health and consumption is higher for (non-mimicking) high-ability individuals than for low-ability individuals.

When cross-border health care takes place, health care to the low-ability type is no longer over-provided to partly accommodate high-ability individuals' demand for shorter waiting times, since they now seek treatment in the other region. However, the self-selection constraint is given larger weight in the optimal provision formula in this case since only the provision to the low-ability individuals must be distorted to relax the constraint.

When individuals are granted the right to cross-border health care (for example by a federal government), but a region wants to induce its citizens not to use this right, health care is over-provided of similar reasons as when cross-border health care is not an option. A difference is that health care in this case is used to relax another self-selection constraint; the constraint that makes high ability individuals preferring to seek treatment in their own region. In this situation the region with the longest waiting time increases their health care expenditures to reduce the waiting time so that the difference in waiting times across regions is small enough in order to prevent cross-border health care.

The result that higher health care expenditures and shorter waiting times can facilitate more redistribution can be contrasted with the result of Hoel and Saether (2003) that longer waiting times can serve as a mean of redistribution. The result of Hoel and Saether (2003) was based on the fact that waiting time makes patients with high waiting costs choose private treatment which reduce the cost of publicly financed health care. The difference in results is due to the mechanism that waiting times affect self-selection in our model. Our results do not contradict those of Besley and Coate (1991) showing that in lack of efficient means of redistribution, like optimal non-linear income taxes, it might be optimal to publicly provide a good of low quality, so that high income household opt for buying the good with higher quality privately instead.

Cross-border health care might be welfare improving for the region with longer waiting times, since it enables residents with high waiting costs (i.e. high-income individuals) to get shorter waiting times than those with lower waiting cost. Since the government can redistribute using non-linear income taxes, cross-border health care might in fact increase welfare for all individuals, but it will, in this model, necessarily increase inequality in health.

The number of patients that seek treatments in another country is currently rapidly increasing in the European Union (Suñol, Garel and Jacquerye, 2009). It is therefore important to increase the knowledge about its causes and consequences. Further research might, for example, analyze the consequences when cross-border health care is driven by returns to scale or other sources of comparative advantages in health care production between regions.

Appendix

Proof of Propositions 1, 2, and 3:

To derive equation (11), move $r\mu_j c'$ in the first order condition for e_j (eq. (10)) to the left hand side and then divide by μ_j. Use $MRS_j^i = \frac{\partial U_j^i/\partial h_j^i}{\partial U_j^i/\partial x_j^i}$ to replace all expression for marginal utilities of health. Then use that equations (6) and (7) imply

$$\gamma_j^1 \omega_j^1 \frac{\partial U_j^1}{\partial x_j^1} = \gamma_j^1 \mu_j + \lambda_j \frac{\partial \widehat{U}_j^2}{\partial x_j^1},$$

and

$$\gamma_j^2 \omega_j^2 \frac{\partial U_j^2}{\partial x_j^2} = \gamma_j^2 \mu_j - \lambda_j \frac{\partial U_j^2}{\partial x_j^2}$$

and rearrange to get equation (11).

Equations (19) and (28) are derived similarly using that equation (14) implies

$$\gamma_{\mathrm{L}}^1 \omega_{\mathrm{L}}^1 \frac{\partial \mathrm{U}_{\mathrm{L}}^1}{\partial \mathrm{x}_{\mathrm{L}}^1} = \gamma_{\mathrm{L}}^1 \mu_{\mathrm{L}} + \lambda_{\mathrm{L}} \frac{\partial \widehat{\mathrm{U}}_{\mathrm{L}}^2}{\partial \mathrm{x}_{\mathrm{L}}^1}$$

and that equation (23) and (24) imply

$$\omega_{\mathrm{L}}^1 \frac{\partial \mathrm{U}_{\mathrm{L}}^1}{\partial \mathrm{x}_{\mathrm{L}}^1} = \mu_{\mathrm{L}}$$

and

$$(\gamma_{\mathrm{L}}^2 \omega_{\mathrm{L}}^2 + \eta_L) \frac{\partial \mathrm{U}_{\mathrm{L}}^2}{\partial \mathrm{x}_{\mathrm{L}}^2} = \gamma_{\mathrm{L}}^2 \mu_{\mathrm{L}} + \eta_L \frac{\partial \mathrm{U}_{\mathrm{L,b}}^2}{\partial \mathrm{x}_{\mathrm{L,b}}^2}.$$

The sign of $\frac{de_j}{d\gamma_j^1}$

The sign of $\frac{de_j}{d\gamma_j^1}$ can be obtained by total differentiating equation (11) with respect to e_j and γ_j^1 which gives

$$\frac{de_j}{d\gamma_j^1} = \left\{ \begin{array}{l} MRS_j^1 - MRS_j^2 + \sum_i \gamma_j^i \frac{\partial MRS_j^i}{\partial \gamma_j^1} + \frac{\partial(\lambda_j/\mu_j)}{\partial \gamma_j^1} \frac{\partial \widehat{U}_j^2}{\partial x_j^1} \left[MRS_j^1 - \widehat{MRS}_j^2\right] \\ + \frac{\partial^2 \widehat{U}_j^2}{\partial (x_j^1)^2} \frac{\partial x_j^1}{\partial \gamma_j^1} \frac{\lambda_j}{\mu_j} \left[MRS_j^1 - \widehat{MRS}_j^2\right] + \frac{\partial \left[MRS_j^1 - \widehat{MRS}_j^2\right]}{\partial \gamma_j^1} \frac{\lambda_j}{\mu_j} \frac{\partial \widehat{U}_j^2}{\partial x_j^1} \end{array} \right\} / (-\sigma) \quad (30)$$

where $\sigma < 0$ is the governments second order condition with respect to e_j. The direct effect is

described by the first two terms in the numerator which are jointly negative as discussed in the text.

As discussed by Aronsson and Blomquist (2008), when the share of low-ability type in a state is increased, the financial burden of the low-ability type is increased. That is, the low-ability likely has to contribute more to finance health care, which they can do either by working more and/or reducing their private consumption, which both makes mimicking less attractive. When mimicking becomes less attractive the high-abilities' tax payment can be increased, which is accomplished by also them working more and/or reducing their private consumption. If health and labor supply are not too strong complements, the effect of reduced consumption dominates the effect on the marginal rate of substitution between health and consumption, implying that the third term is negative.

The fourth term is likely negative since mimicking likely becomes less attractive and the region's budget constraint becomes tighter when γ_j^1 is increased. However, the fifth term is positive given that private consumption of the low-ability type is reduced when γ_j^1 is increased. Lastly, if increased financial burden by the low-ability type primarily affects is marginal rate of substitution by affecting the marginal utility of consumption, the last term in the numerator is negative: the reason is that if the denominator of the MRS-terms are increased proportionally more than the nominators, this reduces the value more of the higher MRS-term.

To conclude, the two direct effects are jointly negative, one indirect effect is positive while three likely is negative given the assumptions made. Without further assumptions we cannot conclude that $\frac{de_j}{d\gamma_j^1} < 0$, but we think that this is highly plausible and it is therefore assumed in the paper.

References

Aronsson T. and Blomquist S. (2008). Redistribution and provision of public goods in an economic federation, Journal of Public Economic Theory 10 (1), 125-143.

Besley T. and Coate S. (1991). Public provision of private goods and the redistribution of income, American Economic Review 81, 979-984.

Blomquist, S. and Christiansen, V. (1995). Public Provision of Private Goods as a Redistributive Device in an Optimum Income Tax Model, Scandinavian Journal of Economics 97, 547-567.

Blomquist, S. and Christiansen, V. (1998). Topping up or Opting out? The Optimal Design of Public Provision Schemes, International Economic Review 39, 399-411.

Boadway, R. and Marchand, M. (1995). The use of Public Expenditures for Redistributive Purposes, Oxford Economic Papers 47, 45-59.

Cullis J.G., Jones P.R. and Propper C. (2000). Waiting lists and medical treatment: analysis and policies, In: Culyer, A.J. and Newhouse, J.P. (eds.) Handbook of Health Economics, Elsevier: Amsterdam.

EU directive 2011/24. 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.

Fossati A. and Levaggi, R. (2008). Delay is not the answer: waiting time in health care and income redistribution. Available at SSRN: http://ssrn.com/abstract=1081928

Gilleskie D.B. (1998). A dynamic stochastic model of medical care use and work absence, Econometrica, 66, 1-45.

Hoel M. and Saether E.M. (2003). Public health care with waiting time: the role of supplementary private health care, Journal of Health Economics 22, 599-616.

Lindsay C.M. and Feigenbaum B. (1984). Rationing by waiting list, American Economic Review 74, 404–417.

Martin S. and Smith P.C. (1999). Rationing by waiting lists: an empirical investigation, Journal of Public Economics 71, 141-164.

Plump J., Redekop W.K., Dekker F.W., vanGeldrop T.R., Haalebos M.M.P., Kingma G.J., Zijlstra F., Tijssen J.G.P. (1999). Death on the waiting list for cardiac surgery in The Netherlands in 1994 and 1995, Heart 81, 593–597.

Siciliani L. and Hurst J. (2005). Tackling excessive waiting times for elective surgery: a comparative analysis of policies in 12 OECD countries, Health Policy 72, 201–215.

Stern N.H. (1982). Optimum Taxation with Errors in Administration, Journal of Public Economics 17, 181-211.

Stiglitz J.E. (1982). Self-selection and Pareto efficient taxation, Journal of Public Economics 17, 213-240.

Suñol, R., Garel, P., and Jacquerye, A. (2009). Cross-border care and healthcare quality improvement in Europe: the MARQuIS research project, Quality and Safety in Health Care 18, i3-i7.

Viscusi W.K. and Evans W.N. (1990). Utility functions that depend on health status: Estimates and economic implications, American Economic Review 80, 353-374.