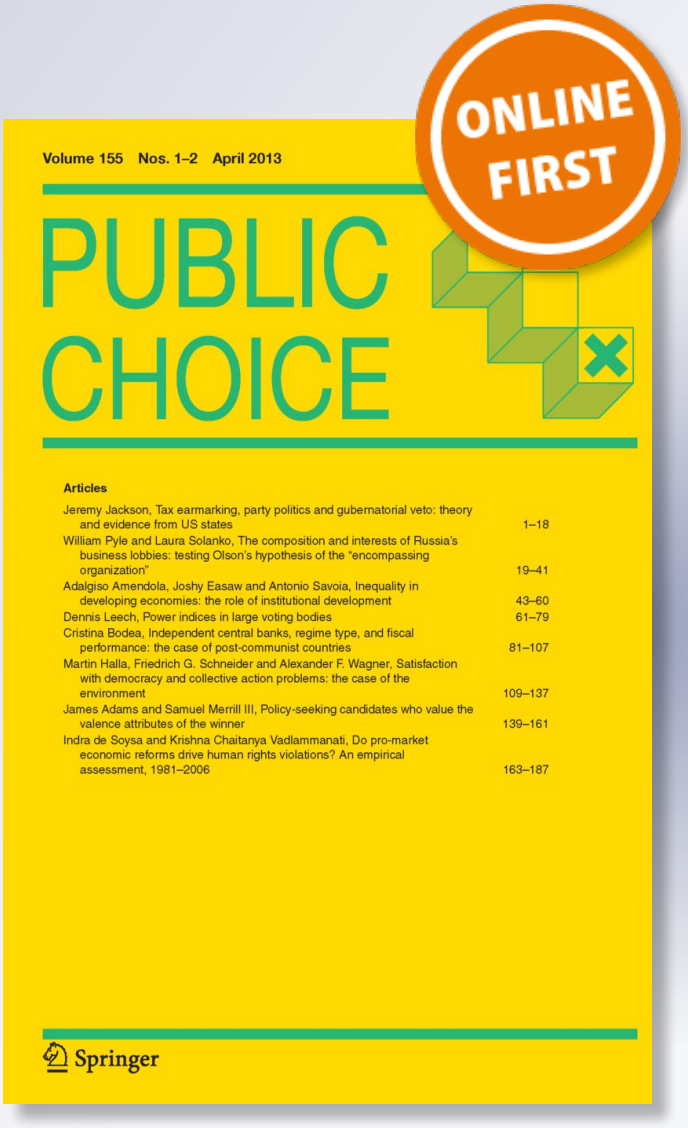
Bureaucratic institutional design: the case of the Italian NHS

Silvia Fedeli, Leone Leonida & Michele Santoni

Public Choice

ISSN 0048-5829

Public Choice
DOI 10.1007/s11127-018-0569-6



Springer

Public Choice
https://doi.org/10.1007/s11127-018-0569-6

Bureaucratic institutional design: the case of the Italian NHS

Silvia Fedeli[1] · **Leone Leonida**[2] · **Michele Santoni**[3]

Received: 1 June 2018 / Accepted: 7 June 2018

Abstract
We propose a model where a regional government's choice of the number of bureaucratic agencies operating in a region depends upon the degree of substitutability and complementarity of the bureaucratic services being demanded. We show that, if the government perceives the citizens' demand as a demand for substitutable services, it will choose provision by two independent agencies. If the government perceives the citizens' demand as a demand for complementary services, it will choose provision by a single consolidated agency. Exogenous shocks to the number of citizens amplify these incentives. Evidence from the Italian National Health Service (NHS) supports this hypothesis. Results show a positive effect of proxies of substitutable services on the number of regional local health authorities and a negative effect of proxies of complementary services. The major immigration amnesties, taken as shocks to the number of citizens entitled to the service, magnify these effects.

Keywords Bureaucratic institutional design · Public local health authorities · Consolidation and decentralization of local health authorities · Italian NHS

JEL Classification D73 · H75 · I18 · L32

> Would you say it is normal to have regions with 7 provinces and 22 Local Health Authorities? […] In my opinion, the idea of regions with 7 provinces and 22

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11127-018-0569-6) contains supplementary material, which is available to authorized users.

✉ Leone Leonida
leone.leonida@kcl.ac.uk

Silvia Fedeli
silvia.fedeli@uniroma1.it

Michele Santoni
michele.santoni@unimi.it

[1] Dipartimento di Economia e Diritto, Facoltà di Economia, Sapienza - Università di Roma, Via del Castro Laurenziano 9, 00161 Rome, Italy

[2] King's Business School, King's College of London, Bush House, 30 Aldwich, WC2B 4BG London, UK

[3] Dipartimento di Economia, Management e Metodi Quantitativi, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milan, MI, Italy

> LHAs is an aberration. So if in agreement with the regions, since this task falls within their responsibilities, we can at last cut the number of managerial positions in the LHAs and apply 'standard costs' throughout, so that the proverbial needle has to cost the same in Calabria as in Lombardy, I would say this is a good thing. (Matteo Renzi, Italian PM, 10th April 2015, http://www.governo.it/media/consiglio-dei-ministri-n58/619, 5:25-6:20; our translation)

1 Introduction

We study the determinants of Italian regional governments' choices concerning the number of local health authorities (LHAs). Our argument is that, *other things being equal*, such decisions depend on whether the government perceives the citizens' healthcare demand as a demand for substitutable or complementary services. We cite empirical results from the Italian National Health Service (NHS) in support of our hypothesis.

Consolidation of local bureaucratic organizations has occurred in different institutional contexts and sectors since the mid-1990 s (e.g., healthcare: Talbot and Johnson 2007 for the United Kingdom; McDaid et al. (2009) for Ireland; local governments: Andrews 2015 for vertical consolidation in England; Blom-Hansen et al. 2016 for horizontal consolidation in Denmark). In Italy, consolidation of the healthcare sector has been on a large scale, the number of LHAs being cut from 638 in 1982 to 143 in 2012. Why do governments consolidate public bureaus and, more specifically, LHAs? The main motivation for consolidation presumably is cost reduction, thanks to economies of scale and scope, lower management and back-up costs, the introduction of cost-effective management styles, and the gain in monopsony power in negotiating with private sector providers (Garside 1999; Fulop et al. 2002 for the United Kingdom; France et al. 2005, p. S190 for Italy). The idea that cutting the number of LHAs can reduce bureaucratic costs, heighten efficiency and free public resources for investment in better healthcare services is popular among Italian national politicians (see the opening quotation). The academic literature also suggests political and institutional motivations for bureaucratic reorganization. Governments interested in policy reforms may see consolidation or decentralization as a way of strengthening their control over the bureaucracy (Lester et al. 1983 on environmental policy in the US states; Del Vecchio and Cuccurullo 2013 on the Italian NHS). Majoritarian electoral systems may facilitate the drive for bureaucratic reorganization (Pollitt 2007 for the United Kingdom). Consolidation/decentralization attitudes, moreover, are related to partisanship. For example, in Britain the Conservative governments of the early 1990 s expanded the number of LHAs in order to mimic competition (Le Grand 1999), whereas the New Labor governments, implementing command and control policies, preferred to have fewer and larger LHAs (Goodwin 2000). Electoral considerations likewise may play a role. For example, hospital mergers in the English NHS under New Labor were less common in election years, especially in swing districts (Gaynor et al. 2012).

In this paper, we propose an additional determinant of a government's decision to consolidate public bureaus, which is driven by strategic incentives to reduce bureaucratic slack. We model the interaction between the government and the bureaus as a two-stage game (see also Fedeli and Santoni 2006). In the first stage, the government chooses whether to consolidate or to keep two bureaus separate. In the second stage, for given amounts of public resources, the government and the bureaus play a Nash budgetary game by choosing

their compliance levels (Miller 1977), namely, respectively, the share of resources to be allocated to the bureaucratic budget and the share of the budget to be devoted to service production. It turns out that in subgame perfect symmetric Nash equilibrium the government can reduce bureaucratic slack by either promoting competition between bureaus when they produce services that it perceives as substitutes, or by consolidating bureaus producing complements. The basic mechanism at work is an externality in bureaucratic compliance levels (i.e., effort levels): provided that the bureaus "care about" the government's demand, their payoffs become interdependent in that an increase in compliance levels in one bureau induces the other bureau to raise (respectively, lower) its own compliance efforts when the services are complements (respectively, substitutes). With complements, the government has an incentive to induce the bureaus to internalize the externality by consolidating them into a single agency, whereas it prefers to keep them separated in the case of substitutes. The model also shows that those incentives are reinforced by shocks to the government's demand for bureaucratic output that were not foreseeable at the time the government designs bureaucratic organization.

We test our hypotheses using regional data on the Italian NHS from 1982 to 2012. The NHS was established in 1978 and has been operative since 1980. It is a universal public healthcare system providing comprehensive insurance coverage and uniform health benefits. It is financed partly by general taxation, along with user copayments for some services. Central and regional governments share responsibility for funding and service provision. On the expenditure side, each of the 20 regional Italian governments sets the level of regional healthcare services provided and determines their administration and organization. Those decisions are subject to a common national legal and regulatory framework, which specifies healthcare standards (termed "essential levels of care", or LEAs). On the financing side, until 2012 the central government set regional healthcare budgets using general revenues and allocating resources on the basis of historical expenditure and healthcare needs, irrespective of regional fiscal capacity (France et al. 2005, p. S194; Francese and Romanelli 2011, p. 6). Major reforms of NHS governance were enacted in 1992 and 1999 to shift financing responsibilities to the regional governments and keep the NHS budget under control, but with mixed results.[1]

In the Italian NHS, each regional government establishes the number of regional LHAs, called *Aziende Sanitarie Locali*. They are independent organizations with full autonomy regarding their legal, organizational, administrative, financial, accounting, managerial and technical responsibilities. The president of the region appoints each LHA's director on the basis of five-year renewable private contracts. The LHAs provide care both directly (through LHA-managed hospitals and territorial agencies offering primary care, outpatient facilities, and other services) and indirectly (paying accredited public and private providers, such as hospitals, nursing homes and laboratories). Each director bargains with the regional government over the LHA's budget. Hence, a rational regional government can organize healthcare services in order to influence the outcome of its negotiations with the LHA's management over the allocation of largely predetermined resources.

The empirical results offer support for the model we propose. As long as each regional government's willingness to pay for healthcare services—which we assume reflects citizens' preferences—is unobservable over time, we proxy that willingness on the basis of society's

[1] Until the mid-1990 s, central government funding amounted to 90% of the overall NHS budget. In 2009 it averaged 49%, with wide variations among the ordinary-statute regions (Mapelli 2012, pp. 107 and 111). Online Appendix 1 outlines the major features of the Italian NHS.

evaluation of physical healthcare resources—inputs used in the transformation process—and produced outputs (Jacobs et al. 2006, chapter 2). In addition, we use the ratio of the number of general practitioners to NHS specialists and the ratio of the number of NHS specialists to hospital beds as proxies for substitutable healthcare services. The idea is that the government perceives the demand for primary care services as a substitute for specialty and inpatient care if, say, increasing minor surgery procedures in general practice reduces patients' demands for in hospital-based minor surgery (Scott 1996). That would also be the case if the government observes that an increase in the number of visits to primary care physicians is associated with a reduction in the number of visits to public healthcare specialists (Atella and Deb 2008). Similarly, the government may perceive NHS specialists and hospital beds to be substitutes, if an increase in the number of privately paid visits to public hospital-based specialists (performing examinations requiring medical equipment, say) lowers the demand for post-procedure recovery in NHS hospitals (Turchetti 2009, pp. 116–117). On the other hand, we take the ratio between the number of general practitioners and NHS hospital beds as a proxy for the demand for complementary healthcare services. That may be the case if the government perceives that the citizens' demand for the detection of illnesses is a demand for both primary care and inpatient care (Fortney et al. 2005). For example, the detection of diabetes mellitus may require both blood pressure measurement by a general practitioner and glycated hemoglobin measurement and visit by a specialist diabetes team.

It turns out that variables proxying for the substitutable (complementary) services show a positive (negative) effect on the number of LHAs. Moreover, demand shocks, proxied by the two major Italian immigration amnesties in 1998 and 2002, magnify the incentives for consolidation or decentralization. The conclusions are not altered by controlling for: common time-specific shocks (supply-side and policy shocks to national healthcare laws and regulations); cost-side determinants (economies of scale); political and institutional factors (including election years, the regional government's political "color", and the type of regional electoral system); additional demand-side determinants of healthcare services (including regional per capita GDP and demographics); and time-invariant unobservable region-specific components (capturing, say, cultural differences).

Our paper also contributes to the empirical literature on the Italian NHS. However, whereas the existing literature focuses on healthcare expenditures and funding (Cellini et al. 2000; Giannoni and Hitiris 2002; Levaggi and Zanola 2003; Bordignon and Turati 2009; Francese and Romanelli 2011; Atella et al. 2014; Fedeli 2015), this paper is the first to model the determinants of the number of LHAs.

The paper proceeds as follows. Section 2 presents a stylized model capturing a regional government's strategic incentives to design the structures of public bureaus. Section 3 presents the data and the empirical model. Section 4 presents the empirical results. Section 5 concludes.

2 The model

Borrowing both from the standard approach to the economic theory of bureaucracies (see, among others, Niskanen 1971; Migué and Bélanger 1974; Breton and Wintrobe 1975, 1982; Forte and Powers 1994), and from the public choice literature's insight that politicians can increase bureaucratic efficiency by properly designing the organizational form of bureaucratic supply (Moe 1984, 2012; Bagnoli and McKee 1991; Ting 2002; Janssen et al. 2003; Bates and Santerre 2008; Bates et al. 2011), and from Horn and Wolinsky (1988),

who develop the idea that consolidation depends on the extent of input substitutability, this section extends Fedeli and Santoni (2006) to the case of demand uncertainty. Consider a regional government willing to offer bureaucratic output (e.g., healthcare services). Output can be produced either separately by two independent agencies or jointly by a single consolidated agency. In the case we examine, agencies are LHAs supplying primary care (such as general practitioner visits and preventive care), secondary care (e.g., specialist visits in hospitals, hospitalization and medical assistance), or both.

The game sequence is as follows. At stage one, the regional government chooses whether to consolidate or to keep two bureaus separate. As long as each regional government has sole responsibility for determining the number of LHAs, this assumption fits the Italian NHS. At stage two, the government negotiates over the budget with the bureau(s). Following the game-theoretic approach of Nikaidô and Isoda (1955) as applied by Miller (1977) and Fedeli (1999) to compliance relationships inside bureaucracies, each agent chooses simultaneously its own Nash compliance level. In particular, the government sets the share of public resources it allocates to the bureaus as a budget and the bureaus set the share of their budget that actually is used for producing the desired output. Depending on the outcome of stage one of the game, two subgames are considered. In the consolidation subgame, the government and the single bureau choose their own compliance strategies, by taking as given the strategies simultaneously chosen by the other agent. In the separation subgame, simultaneous bilateral Nash budget games are played between the government and each single bureau. Similar to Horn and Wolinsky's (1988) Nash-in-Nash bargaining solution,[2] each bilateral negotiation selects its Nash equilibrium under the assumption that a Nash equilibrium will occur in the other negotiation as well. Then, the equilibrium bureaucratic organizational form and compliance levels arise in sub-game perfect Nash equilibrium.

The government's preferences are

$$G = \left\{ \alpha\left[\mathbf{Q_1}+\mathbf{Q_2}\right] - \left[\frac{\beta\left(\mathbf{Q_1^2}+\mathbf{Q_2^2}\right)}{2}\right] - \mathbf{2}\gamma Q_1 Q_2 \right\} + \sum_{k=1}^{2}(R - B_k), \tag{1}$$

The first term represents a quadratic evaluation function of the bureaucratic production of differentiated outputs Q_k, $k=1,2$, with $\alpha>0$, $\beta>0$, $\beta^2>4\gamma^2$ and $\beta>2|\gamma|$ (Singh and Vives 1984). We assume that the function derives from the preferences of the electoral constituency (citizens or potential patients). The second term is the political rents the government obtains from the budgetary process, which is equal to the difference between public resources R and the budget $B_k=Rg_k$ the regional government assigns to the bureaus, where $g_k \in [0,1]$ is the share of resources it allocates to the production of output $k=1, 2$. R is predetermined by the central government based on observables (e.g., demographics), an assumption that fits the Italian NHS. Governmental maximization of (1) with respect to Q_k gives

$$V_k = \alpha + \varepsilon - \beta Q_k - 2\gamma Q_l, \tag{2}$$

[2] Gaynor et al. (2015, pp. 254–258) apply the Horn and Wolinsky (1988) model to hospital-insurer price negotiations. Inhe current paper, decisions to merge bureaus or not are taken by the government, not by producers. The Horn and Wolinsky (1988) solution concept originally was developed to examine incentives for horizontal mergers between trade unions in the presence of an exclusive vertical relationship with a single employer. Collard-Wexler et al. (2018) review the industrial organization and labor economics literatures adopting the solution and provide microeconomic foundations for it.

where we enter a common additive shock ε with $E(\varepsilon)=0$ and $E(\varepsilon^2)=s^2$, to the solution. V_k, $k=\{1,2\}$ and $k \neq l$, represents the government's perception of the citizens' willingness to pay for Q_k, given Q_l. For $\gamma<0$ the two outputs are complements. In that case, the willingness to pay for, say, general practitioners' diagnostic services and referral to hospital increases with the demand for hospitalization. For $\gamma>0$ the outputs are substitutes. As such, the willingness to pay for, say, specialist visits in hospital declines when the demand for primary care physician visits rises. For $\gamma=0$, the two outputs are independent. The shock ε affects linearly the government's reservation price, thus the willingness to pay for bureaucratic output. That term can be interpreted as a shock to the population that is eligible, e.g., for public healthcare coverage. The shock occurs after the government's organizational choice, but its realized value is public knowledge before stage two takes place. Regarding the Italian NHS, the shock can be related to the inflows and outflows of immigrant workers and their families in each region associated with national immigration amnesties. Following Klemperer and Meyer (1986), assume that the size of the shock is small enough so that both prices and bureaucratic output levels are positive.

Turning to bureaucratic preferences, with two independent bureaus (LHAs) we have decentralization, D:

$$LHA\ D_k = V_k Q_k + B_k(1 - h_k) \tag{3}$$

With one consolidated bureau we have centralization, C:

$$LHA\ C = \sum_{k=1}^{2} LHA \quad D_k = \sum_{i=k}^{2} V_k Q_k + \sum_{i=k}^{2} B_k(1 - h_k) \tag{4}$$

Equations (3) and (4) assume that public agencies care about bureaucratic revenue and evaluate their activities according to V_k, namely the government's perception of the citizens' marginal willingness to pay for good $k=1, 2$. The latter assumption (originating from Niskanen 1971) fits the Italian NHS case, as the president of the region appoints the directors of the LHA, which means that they cannot ignore the politicians' wishes. However, bureaus also are interested in slack. Slack is represented by the second term of both equations, where $h_k \in [0, 1]$ denotes the share of the public budget B_k that the bureau allocates to production. For simplicity, we assume that production occurs with constant-returns-to-labor technology at the minimum (symmetric) production cost $c>0$, with $0<c<\alpha$. Hence, total production costs are $TC_k=cQ_k$. That assumption enables us to single out the government's strategic incentives in designing bureaucratic structure that are related to the demand side. In the empirical section, we control for possible determinants of cost, including the returns-to-scale regime. Note that, although production occurs at minimum average cost, the budgetary process generates ex post inefficiencies as far as some slack arises in equilibrium.

Solving the model by backward induction, in stage two the government and the bureaus choose compliance levels, given the centralization/decentralization choice at stage one and the realized value of the shock ε. From previous assumptions, public resources for producing good k can be written as $h_k g_k R=cQ_k$, implying that $Q_k=h_k g_k R/c$, for $k=1,2$. Substituting this back into Eqs. (1) and (3) or (4), preferences can be formulated in terms of the budgetary strategies g_k and h_k (Miller 1977) and the value of the shock ε. In the separation subgame D, the government engages in a simultaneous compliance game with each bureau. The government chooses g_k by maximizing Eq. (1), taking as given the budgetary strategy of the bureau choosing h_k to maximize Eq. (3) given g_k. Both agents take the Nash outcome of the parallel compliance

game between the government and the other bureau as given, yielding the symmetric Nash equilibrium unconsolidated compliance levels:

$$\hat{g}_1^D = \hat{g}_2^D = \frac{(\alpha - c + \varepsilon)[\alpha\beta + (\beta + 2\gamma)(c - \varepsilon)]}{4R(\beta + \gamma)^2},$$
$$\hat{h}_1^D = \hat{h}_2^D = \frac{2c(\beta + \gamma)}{[\alpha\beta + (\beta + 2\gamma)(c - \varepsilon)]}$$

In the consolidation subgame C, the government sets simultaneously g_k and g_l by taking as given the strategies h_k and h_l simultaneously set by the single bureau to maximize Eq. (4) for given *gs*. The solution yields the Nash equilibrium compliance levels:

$$\hat{g}_1^C = \hat{g}_2^C = \frac{\left[\alpha^2 - (\varepsilon - c)^2\right]}{4R(\beta + 2\gamma)},$$
$$\hat{h}_1^C = \hat{h}_2^C = \frac{2c}{[\alpha + c - \varepsilon]}.$$

Note that compliance levels always are positive and fall short of unity, provided the size of the shock is small enough. That result implies positive political rents and bureaucratic slack in Nash equilibrium. At stage one of the game, the government knows the distribution of the shock and its own state-contingent payoffs at stage two. Computing the government's expected value function under D or C yields

$$E\left(\hat{G}^D\right) = 2R + \frac{(\alpha - c)^2(\beta + 2\gamma)}{4(\beta + \gamma)^2} + \frac{s^2(\beta + 2\gamma)}{4(\beta + \gamma)^2}$$
$$E\left(\hat{G}^C\right) = 2R + \frac{(\alpha - c)^2}{4(\beta + 2\gamma)} + \frac{s^2}{4(\beta + 2\gamma)}, \tag{5}$$

Equation (5) shows that additive uncertainty raises the government's expected payoff above its deterministic level by a term that is proportional to the variance of the shock s^2. Hence, the government chooses centralization as its subgame perfect Nash equilibrium strategy if and only if

$$E\left(\hat{G}^C\right) - E\left(\hat{G}^D\right) = \gamma\left[(\alpha - c)^2 + s^2\right]\left[\frac{(2\beta + 3\gamma)}{4(\beta + \gamma)^2(\beta + 2\gamma)}\right] > 0, \tag{6}$$

namely for $\gamma < 0$, meaning that bureaus produce complementary outputs from the electorate's viewpoint. The intuition for this result is as follows. Provided that bureaus evaluate outputs according to the government's demand, their payoffs become interdependent. When the outputs are complements, an increase in compliance by one bureau raises (lowers) the marginal utility of higher compliance by the other bureau. The government therefore has a strategic incentive to induce bureaus to internalize that externality by consolidating them into a single agency at stage one of the game, as long as consolidation reduces slack and raises the government's expected utility. When the outputs are substitutes, $\gamma > 0$, however, the externality takes the opposite sign, and a single consolidated agency would set a lower compliance level, resulting in more slack in Nash equilibrium. In that situation, the government has a strategic incentive to keep the two bureaus separated. Hence, unlike in Bagnoli and McKee (1991) and contrary to the received view, promoting competition among

bureaus does not necessarily increase bureaucratic efficiency, as that outcome depends on the nature of the externality in the budgetary process. Moreover, reducing bureaucratic slack in the presence of complements also implies an increase in government's compliance (as long as the government's marginal utility of greater compliance increases the bureau's compliance level, i.e., g_i and h_i are strategic complements from the government's viewpoint). As a consequence, bureaucratic production is higher when complements are produced by a consolidated bureau (i.e., $Q_i^C > Q_i^D$ for $\gamma < 0$). Moreover, uncertainty enhances the government incentives to consolidate bureaus with complements and to keep them separated with substitutes.

2.1 The stochastic solution

In order to evaluate more explicitly the difference between the stochastic and the deterministic equilibrium solutions of the model, it is useful to decompose the stochastic component of the government's expected value function under centralization C and rewrite Eq. (5), as follows:

$$\frac{s^2}{4(\beta+2\gamma)} = \underbrace{\frac{s^2}{4(\beta+2\gamma)}}_{\text{stochastic gains from rents}} = \underbrace{\frac{s^2}{4(\beta+2\gamma)}}_{\text{stochasticloss from production}} \tag{7}$$

Equation (7) shows that the stochastic shock alters the government's tradeoff between the utility from political rents and that from the outputs. Let us explicate the economic intuition behind this result. In the stochastic case, the budget provided to the consolidated bureau is smaller, namely the government's expected compliance $E\left(\widehat{g}^C\right) - = \frac{\alpha^2 - c^2 - s^2}{4R(\beta+2\gamma)}$ is lower than in the deterministic case, $g^C = \frac{\alpha^2 - c^2}{4R(\beta+2\gamma)}$, implying that expected political rents are higher in the stochastic than in the deterministic case. However, although the equilibrium level of the outputs is the same in both cases, $E\left(\frac{\widehat{g}^C \hat{h}^C R}{c}\right) = E\left(\frac{\alpha+\varepsilon-c}{2(\beta+2\gamma)}\right) = \frac{\alpha-c}{2(\beta+2\gamma)}$, by Jensen's inequality the expected indirect utility from the outputs is lower in the stochastic than in the deterministic case, as long as the following is satisfied

$$E\left[\left(\frac{\widehat{g}^C \hat{h}^C R}{c}\right)^2\right] = \left[\frac{(\alpha-c)^2 + s^2}{4(\beta+2\gamma)^2}\right] > \left[E\left(\frac{\widehat{g}^C \widehat{h}^C R}{c}\right)\right]^2 = \left(\frac{\alpha-c}{2(\beta+2\gamma)}\right)^2.$$

Considering both effects and given that the government's payoff is concave in outputs, the larger expected political rents in the stochastic case more than offset the reduction in the expected gains from the outputs. As a result, the expected payoff is increasing in the variance of the shock. The analysis of the stochastic component of the government's expected value function under decentralization D is similar. It follows that the latter also will be increasing in the variance of the shock. Hence, uncertainty will enhance the government's incentive to choose C with complements and to choose D with substitutes.

2.2 Exogenous shocks to the demand slopes

Following Klemperer and Meyer (1986), assume now that the exogenous shock influences the slopes of demand

$$V_k = \alpha - \frac{\beta}{\varepsilon} Q_k - \frac{2\gamma}{\varepsilon} Q_l, \tag{8}$$

with $k,\ l=1,2,\ k\neq l$ and with $E(\varepsilon)=1$, $E(1/\varepsilon)>1$, $E(\varepsilon^2)=s^2>1$. Such a shock leaves both the degree of substitutability between bureaucratic outputs and the reservation price unaffected. The shock can be interpreted as capturing as well an unexpected increase in the local population eligible for accessing bureaucratic output. Solving the model by backward induction, the players' budgetary strategies at a symmetric Nash equilibrium when the government deals with one single bureau are

$$\begin{aligned} \breve{g}_1^C &= \breve{g}_2^C = \frac{(\alpha - c)[2\alpha - (\alpha - c)\varepsilon]\varepsilon}{4R(\beta + 2\gamma)}, \\ \breve{h}_1^C &= \breve{h}_2^C = \frac{2c}{2\alpha - (\alpha - c)\varepsilon}. \end{aligned} \tag{9}$$

Therefore, at stage one of the game, the government's expected payoff is

$$E\left(G^C\right) = \underbrace{2R}_{\text{deterministic payoff}} + \underbrace{\frac{2s^2(\alpha - c)^2}{4(\beta + 2\gamma)}}_{\text{stochastic gains in terms of rents}} - \underbrace{\frac{s^2(\alpha - c)^2}{4(\beta + 2\gamma)}}_{\text{stochastic loss in terms of output evaluation}}. \tag{10}$$

The symmetric budgetary strategies and the government's expected payoff when dealing with two bureaus are:

$$\begin{aligned} g_1^D = g_2^D &= \frac{(\alpha - c)[2\alpha(\beta + \gamma) - (\alpha - c)(\beta + 2\gamma)\varepsilon]\varepsilon}{4R(\beta + \gamma)^2}, \\ h_1^D = h_2^D &= \frac{2c(\beta + \gamma)}{2\alpha(\beta + \gamma) - (\alpha - c)(\beta + 2\gamma)\varepsilon}. \end{aligned} \tag{11}$$

$$E\left(G^D\right) = \underbrace{2R}_{\text{deterministic payoff}} + \underbrace{\frac{2s^2(\alpha - c)^2(\beta + 2\gamma)}{4(\beta + 2\gamma)}}_{\text{stochastic gain in terms of rents}} - \underbrace{\frac{s^2(\alpha - c)^2(\beta + 2\gamma)}{4(\beta + 2\gamma)}}_{\text{stochastic loss in terms of output evaluation}}. \tag{12}$$

Equations (10) and (12) show that the government's expected payoffs are increasing in the variance of the shock, yielding:

$$E\left(G^C\right) - E\left(G^D\right) = -\gamma\left[(\alpha - c)^2 s^2\right]\left[\frac{2\beta + 3\gamma}{4(\beta + 2\gamma)(\beta + \gamma)^2}\right] > 0. \tag{13}$$

Uncertainty about the elasticity of the demand curve reinforces the deterministic incentives of the government to consolidate the bureaus with complements ($\gamma<0$) and

to keep the bureaus separated with substitutes ($\gamma > 0$). This confirms the previous findings derived under the assumption of an additive exogenous shock.

3 Data description and empirical framework

Generalizing the implications of the theoretical model, we derive two testable predictions.

H1 There is a negative (positive) correlation between the number of regional bureaucratic agencies, e.g., LHAs, and the citizens' demand for complementary (substitutable) healthcare services.

H2 Shocks to the demand for bureaucratic output strengthen the negative (positive) correlation between the number of bureaus and the electorate's demand for complementary (substitutable) healthcare services.

3.1 Data

The empirical analysis builds on yearly data for the 20 Italian regions over the 1982–2012 period. LHAs or *Aziende Sanitarie Locali*, our dependent variable, is the number of basic independent agencies of the Italian NHS that provide comprehensive healthcare services. Each LHA receives its budget from the regional government. National law empowers the regions to determine the number and organization of the LHAs (Law 833/1978, Law 502/1992, and Legislative Decree 228/1999). Over the sample period, all 20 regions reduced to some extent the number of LHAs. The 1992 NHS reform actually triggered LHA consolidation in almost all of the regions in 1995, though to different extents and at a different pace across regions. (Del Vecchio and Cuccurullo 2013, pp. 36–37, provide a taxonomy of regional consolidations in 1995–2013.)

To test the model's predictions, we need variables measuring the regional governments' perception of citizens' demand for healthcare services. Because regional citizens' willingness to pay is unobservable, as in Jacobs et al. (2006), we proxy willingness to pay using the social evaluation of physical inputs used in the transformation process and produced healthcare outputs. The first variable we consider is the ratio of the number of general practitioners to the number of specialists working for public and private accredited NHS hospitals per 1000 inhabitants (*GPs/Specialists*), which we take as a proxy for the demand for substitutable services. That interpretation is consistent with theory and evidence. For example, Atella and Deb (2008) use survey data to show that, when unobserved heterogeneity is taken into account, patients perceive GPs and in-hospital specialists as substitute medical care providers. Fortney et al. (2005, pp. 1424–1425) argue that from the patient's standpoint primary care is a substitute for secondary care when by prevention or early detection of disease (e.g., "prevention of stroke by treatment of hypertension") or by management of chronic conditions (e.g., "control of blood sugar to avert kidney failure in patients with diabetes mellitus"), the need for specialty care is delayed or avoided. Evidence of substitutability is drawn from the US Department of Veteran Affairs (Fortney et al. 2005). Wright and Ricketts (2010) show that a greater concentration of primary care

physicians, depending on the geographical level of the data, is associated with a reduction in inpatient hospital admissions and emergency room visits (also see Scott 1996; van Dijk et al. 2014).

Our second proxy for substitutability is the ratio of the number of specialists to the number of accredited NHS hospital beds per 1000 inhabitants (*Specialists/Hospital beds*). The number of specialists is a proxy for healthcare demand by fee-paying patients and the number of beds is taken as a proxy for their demand for hospitalization services. That interpretation is based on the argument that, in Italy, specialists working in NHS hospitals for fixed salaries also can see patients as private professionals outside their contracted hours of public sector work. Private professional activities can be offered either *intramoenia* (in the specialist's NHS hospital) or *extramoenia* (outside it). The overwhelming majority (more than 94%) operate *intramoenia*, which may be explained by the incentive mechanisms built into the law. When visiting outpatients and inpatients *intramoenia*, specialists use NHS facilities, so each hospital autonomously fixes the fees that the specialists' patients must pay, subject to regional regulations. It is widely believed that private specialists' visits reduce the need for NHS hospital recovery by providing diagnosis and diagnostic tests that otherwise would be part of the NHS hospitalization service package (Turchetti 2009, pp. 116–117). Note that, although the choice between the *intramoenia* and *extramoenia* regimes was introduced in 1996 (Law 662/1996, Legislative Decree 229/1999, and Laws 138/2004 and 120/2007), since 1980 the specialists working in public hospitals have had the right to private professional practice as well (Presidential Decree 761/1979).

Our third empirical proxy captures the demand for complementary healthcare services. In Italy, as in other countries such as the United Kingdom, the GP's prescription determines whether a patient needs publicly provided specialty care, meaning that patients can see the GPs' services of "gatekeeping" and referral as complementary to the hospitalization services (Maio and Manzoli 2002). We take the number of beds in public and private accredited NHS hospitals as a proxy for the citizens' demand for hospitalization services and proxy the demand for GPs' gatekeeping services with the number of GPs (van Dijk et al. 2014). Thus, the ratio of the number of GPs to the number of NHS hospital beds per 1000 inhabitants (*GPs/Hospital beds*) serves as a measure of the demand for complementary healthcare services.

Summary statistics for the variables used and data sources are reported in Table 1.

3.2 The empirical model

We assume that the variation in the number of LHAs in region i at time t, ΔLHA_{it}, is given by:

$$\begin{aligned}\Delta LHA_{it} =& \rho LHA_{it-1} + \pi_1(GPs/Specialists)_{it-1} + \pi_2(GPs/Hospital\ beds)_{it-1} \\ &+ \pi_3(Specialists/Hospital\ beds)_{it-1} + \boldsymbol{\theta}\mathbf{X}_{it-1} + \mu_t + u_{it},\end{aligned} \tag{14}$$

where $GPs/Specialists_{it-1}$, $GPs/Hospital\ beds_{it-1}$ and $Specialists/Hospital\ beds_{it-1}$ are the proxies for the regional demand for differentiated healthcare services; μ_t and u_{it} are time dummies and the random *iid* error term.

Equation (14) is a convenient representation of the adjustment of the number of LHAs towards its long-run level. The dependent variable is the variation in the observed number of regional LHAs, which is zero until the regional government decides otherwise. There are 55 such changes in our sample period, all of them representing consolidation. Hence, the dependent variable is censored, which makes the Tobit approach the natural estimation

Table 1 Variables, sources and summary statistics

Variable	Definition	Obs.	Mean	SD	Min	Max
LHA	Number of regional LHAs (Aziende Sanitarie LocaK)	620	18.61	19.32	1.00	84.00
HT	Number of regional Hospital Trusts (Aziende Sanitarie Ospedaliere)	620	2.64	5.37	0.00	29.00
Specialists/Hospital beds	Ratio between the number of specialists and the number of beds in public and private accredited NHS hospitals	580	0.36	0.16	0.11	0.75
GPs/Hospital beds*	Ratio between the number of General Practitioners (GPs) and the number of beds in public and private accredited NHS hospitals	600	0.16	0.06	0.03	0.29
GPs/Specialists*	Ratio between the number of GPs and the number of specialists working in public and private accredited NHS hospitals	580	0.48	0.14	0.19	0.99
Economies of scab*	Ratio between the number of beds in public and private accredited NHS hospitals and the number of hospitals	600	225.49	89.39	92.39	628.00
Population % population > 65	Regional population	620	2878.14	2275.99	112.35	9900.00
	Percentage share of the regional population aged over 65	620	17.75	3.77	9.00	28.00
Mortality	Regional mortality rate	600	99.58	14.65	73.00	145.00
Infant mortality	Regional infant mortality rate	600	61.63	31.33	2.00	183.00
Left wing governing**	Dummy = 1 for years with center-left regional governments (including governments with PCI-PSI, DC-PSDI-PSI-PRI-PLI and DC-SVP coalitions in the years 1982–1993) and 0 otherwise	620	0.75	0.43	0	1
Election**	Dummy = 1 for years with general political ebction or regional ebction and 0 otherwise	620	0.43	0.50	0.00	1.00
Change of electoral rule**	Dummy = 1 under the plurality system and 0 under the proportional system The regional electoral system changed from proportional to mixed-pluralitv in 1994	620	0.53	0.50	0	1
Direct ebction of the president**	Dummy = 1 since the introduction of direct ebction of the president of the region. In ordinary-statute regions it takes value 0 up to 1998; in special-statute regions it takes value 0 up to 2000	617	0.43	0.50	0	1
Special-statute regions**	Dummy = 1 for special-statute regions and 0 otherwise	620	0.25	0.43	0	1
Turco Napolitano**	Variation in the number of regional population due to new residence permits (1998–2001)	620	7797	32,534	0	312,254
BossiFini**	Variation in the number of regional population due to new residence permits (2002–2011)	620	39,788	106,242	0	940,740

Period 1982–2012. Yearly frequency. *Sources* ISTAT (Health for all dataset); *Ministero deDa Programmazione Economica, 1982–1997; **Ministero deUTnterno. Special-Statute Regions are: Vale D'Aosta, Trentino-Alto Adige, Friuli Venezia Giulia, Sicily, Sardinia

framework. Our estimation strategy is based on: (i) the nature of the dependent variable, which often is zero; (ii) the assumption that the regional fixed effects are not statistically significant because they are removed by differencing the dependent variable (a hypothesis we can test for); and, finally, (iii) the reasonable thesis that given the "large T, small N" setting, the bias introduced by the presence of the lagged dependent variable is in any case not sizeable (less than 2% even if N were large, which is far from the case here). The independent variables all are assumed to affect the variation in the number of LHAs at time $t-1$, because of the time needed for the regional government to adjust its budget target once they are observed. That assumption also helps to control for simultaneity and reverse causality. Entering the lagged dependent variables in our estimating equation help minimize potential omitted variables bias in the model, and ρ gives a measure of the speed of adjustment to the desired level. Finding ρ to be statistically significant and negative helps rule out non-stationarity of the dependent variable.

The three main independent variables of interest are non-linear combinations of three variables, *GPs*, *Specialists* and *Hospital beds*, allowing us to identify the three parameters in which we are interested. We expect that the higher is the *GPs/Specialists* ratio, the higher is the regional demand for substitutable health care services. Similarly, because private visits reduce the need for hospitalization, patients see specialists and hospital recovery as substitute healthcare service providers. Accordingly, we hypothesize that the higher is the *Specialists/Hospital beds* ratio, the higher is the regional demand for substitute services. Hence, the higher is the *GPs/Specialists* ratio or the higher is the *Specialists/Hospital beds* ratio, the greater is the government's desired number of LHAs. Finally, we expect that the higher is the *GPs/Hospital beds* ratio, the higher is the regional demand for complementary healthcare services and, hence, the lower is the desired number of LHAs.

The vector **X** is a set of control variables drawn from the previous literature that are meant to capture economic, sociodemographic, political and institutional features that potentially affect the regional governments' organizational choices. In what follows, we describe only the control variables of our preferred model (see online Appendix 2 for a detailed description of the estimating model, identification, the selection of the set of regressors, and robustness checks). The first control variable is the (natural logarithm of) average number of beds per NHS hospital, which Bordignon and Turati (2009, p. 313) take as a proxy for economies of scale in the production of healthcare services. Because LHA consolidation often is motivated by taking advantage of scale economies, we expect that the larger is the average number of beds per hospital, the lower is the desired number of regional LHAs. The empirical evidence shows that non-linear economies of scale exist in the Italian hospital sector. As the number of beds per hospital rises above a certain threshold, diseconomies of scale are likely to emerge (Cellini et al. 2000, p. 511; Schiavone 2008, pp. 11–13). If that is the case, an increase in the number of beds per hospital is not necessarily associated with an efficiency incentive for consolidation. In order to check for potential non-linearities, we include as a regressor the squared value of the log of the regional ratio of beds to hospitals.

The second control variable is the number of *Aziende Sanitarie Ospedaliere* (Hospital Trusts, hereafter HTs). The HTs were introduced by the 1992 NHS reform (Article 4 of Law 502/1992; Law 405/2001). The HTs are highly specialized hospitals with a national profile. Their legal status and activities broadly are akin to those of the UK Hospital Trusts. They are independent public enterprises with full legal, administrative and managerial autonomy. They produce healthcare services directly. They serve as reference structures for emergency services and acute care within their geographical areas. The president of the region appoints their directors. Each regional government is empowered to determine the

number of HTs, subject to criteria established by the national law. The HTs receive their budgets from the LHAs, which purchase their services. In practice, such contractual relations exist in four regions (Lombardy, Emilia-Romagna, Tuscany and Umbria). The other 16s regional governments allocate budgets to HTs directly (France et al. 2005, p. S191; Mapelli 2012, p. 220). The introduction of the HTs (most notably in Lombardy, Sicily and Campania) partially offset reductions in the number of LHAs. Finally, we enter a dummy variable to control for electoral cycles (*Election*), which is equal to 1 in the years with a general political election or regional election and is equal to 0 otherwise. We expect that regional governments with partisan/electoral motivations will reorganize their LHAs in closer alignment with citizens' wishes (their healthcare demands) and in proximity of election years.

We identify regional demand shocks by controlling for changes in population associated with the major national immigration amnesties of 1998 and 2002. Interacting those variables with our proxies, the model is:

$$\begin{aligned}\Delta LHA_{it} =& \rho LHA_{it-1} + \left(\pi_1 + \varphi_1 Shock_{it}^{(j)}\right)(GPs/Specialists)_{it-1} + \left(\pi_2 + \varphi_2 Shock_{it}^{(j)}\right)(GPs/Hospital\ beds)_{it-1} \\ &+ \left(\pi_3 + \varphi_1 Shock_{it}^{(j)}\right)(Specialists/Hospital\ beds)_{it-1} + \boldsymbol{\theta}\mathbf{X}_{it-1} + \mu_t + u_{it},\end{aligned} \tag{15}$$

$Shock_{it}^{(j)}$, $j=1,2$ is the change in population following the 1998 and 2002 amnesties, including family reunions. Over the sample period, actually, Italy enacted five immigration amnesties for undocumented workers, namely, in 1986, 1990, 1995, 1998 and 2002. Those of 1998 and 2002 involved the largest numbers (Fasani 2009) and had important implications for the NHS. Specifically, in 1998 a center-left government sponsored the "Turco-Napolitano" Law (Law 40/1998, Legislative Decree 286/1998), named after the two cabinet members who proposed it to parliament. The law laid down Italian policy on immigrants' entry, residence and working conditions and on the deportation and control of undocumented immigrants. The law made it compulsory for documented immigrant workers to join the NHS, with the same tax obligations and healthcare rights as Italian citizens. It specified that NHS coverage had to be extended to their family members with regular residence permits, including their children. It further established that regularized immigrant workers had the right to apply for the reunion with their spouses, children and non-working parents living outside Italy. The amnesty associated with the law regularized 217,124 undocumented workers who already were living in the country. In 2002, a center-right government sponsored the "Bossi-Fini" Law (Law 189/2002), likewise named after two cabinet members. The Bossi-Fini Law amended Turco-Napolitano and introduced new clauses regulating undocumented immigration to Italy. It came into force in August and was followed by Decree Law 195/2002 on the procedures for regularizing undocumented immigrant workers already in the country in September. That amnesty regularized 702,156 undocumented workers, 47% of all those regularized during our sample period. All regularized immigrant workers received residency permits with durations equal to the lengths of their employment contracts. The residence permits gave them the right to access the NHS directly as well as to apply for family reunion (Devillanova 2008; Pellizzari 2013). Based on our theoretical model, we expect φ_1, φ_2 and φ_3 to have the same signs as π_1, π_2 and π_3. If so, the evidence tells in favor of H2. Note that, because the amnesty variables clearly are demand-side regressors, such evidence would further support their use as proxies for the government's perception of citizens' healthcare demands.

Table 2 Results for Hypothesis 1

Model	(a)	(b)	(c)	(d)
Dependent variable	$\Delta LHA_{i,t}$	$\Delta LHA_{i,t}$	$\Delta LHA_{i,t}$	$\Delta LHA_{i,t}$
$LHA_{i,t-1}$	− 0.068***	− 0.051***	− 0.274***	− 0.043***
	(0.022)	(0.016)	(0.062)	(0.007)
Specialists/Hospital $beds_{i,t-1}$	0.117***	0.166***	0.337***	0.251***
	(0.041)	(0.047)	(0.089)	(0.064)
GPs/Hospital $beds_{i,t-1}$	− 0.224**	− 0.304***	− 0.467*	− 0.422***
	(0.094)	(0.102)	(0.262)	(0.138)
GPs/$Specialists_{i,t-1}$	0.069**	0.083***	0.132**	0.086***
	(0.028)	(0.029)	(0.058)	(0.036)
Economies of $scale_{i,t-1}$		0.176**	0.397*	0.126**
		(0.085)	(0.286)	(0.080)
Economies of $scale^2_{i,t-1}$		− 0.016**	− 0.036	− 0.011**
		(0.008)	(0.026)	(0.007)
$\Delta HT_{i,t-1}$		− 1.608***	− 1.552***	− 2.001***
		(0.457)	(0.357)	(0.097)
$Election_{i,t-1}$		1.402**	1.111**	1.098**
		(0.648)	(0.461)	(0.430)
Constant	− 4.219**	− 55.722**	− 121.738	− 43.981**
	(1.821)	(24.267)	(81.682)	(22.229)
F-test (31, 580) for time dummies	8.600***	8.890***	8.450***	47.040***
F-test (38, 510) for region specific time trends			4.210***	
Wald test for exogeneity (p value)				6.61 (0.158)
Log-likelihood	− 1643.52	− 1517.94	− 1475.27	
Pseudo-R^2	0.0696	0.1149	0.1398	
Number of observations	580	560	560	560

Column (a) reports the Tobit estimates for the impact of the proxes for government's perception of citizens' demand. Column (b) reports coefficients controlled for supply and policy shocks, where the non-linear specification of the economies of scale, the variation in the number of hospital trusts, and the presence of an election dummy are added to the set of regressors. Column (c) controls estimated parameters for region-specific time trends and, finally, Column (d) reports results controlled for endogeneity. Standard errors in parenthesis are robust to autocorrelation and heteroschedasticity. ***$p<0.01$, **$p<0.05$, *$p<0.1$

4 Results

The results for the initial specification of the estimating model are given in Column (a) of Table 2. Standard errors are consistent with autocorrelation and heteroscedasticity. The lagged dependent variable is statistically significant and negative, as expected, which supplies evidence in favor of the existence of some inertia in the adjustment process. The coefficient is small, in line with the hypothesis of gradual adjustment of the number of LHAs to the long-run level. The results support H1. They suggest that the change in the number of LHAs is positively correlated at standard levels of significance with *GPs/Specialists* and *Specialists/Hospital beds*, our measures of the government's

perception of citizens' demand for substitutes, and negatively correlated with *GPs/Hospital beds*, our measure for the demand for complements.

The results reported in Table 2 are robust to region-specific fixed effects (removed by first differencing), and to common supply shocks, controlled for with the full set of time dummies. In one second exercise, we add several variables to the set of regressors, following a general-to-specific approach. The results of that approach are summarized in Column (b) of Table 2, where our model in Column (a) is augmented by adding our measure of economies of scale (in its non-linear specification), the variation in the number of HTs, and the dummy for elections. The variable capturing economies of scale is statistically significant at the 5% level with a positive sign, and negative for its squared value. Therefore, an increase in the number of beds per hospital is associated at first with an increase in the desired number of LHAs, but above a certain threshold a further increase gives the government an incentive to consolidate.

The value of the *Economies of scale* variable peaks at around 240–280 beds per hospital, which compares well with the literature on that issue. The results also suggest that the change in the number of HTs is negatively correlated, at the 1% significance level, with the change in the number of LHAs. Our interpretation is that governments perceive their citizens seeing the acute healthcare services provided by HTs as complementary to the primary and secondary healthcare that the LHAs either produce directly or procure from accredited private providers. Therefore, an increase in the number of HTs is taken as an increase in a region's demand for complementary healthcare services, which in our model should be negatively correlated with the number of LHAs. Finally, we control for whether an election was held in the year prior to the change in the number of LHAs. The sign of *Election* suggests that the number of LHAs is more likely to change in years preceding the elections.

Concerns relate to the potential omitted variable, to potential correlation of the error term with the beginning-of-period number of *LHAs*, LHA_{it-1}, and to potential correlation of the former with the main independent variables of interest, namely $GPs/Specialists_{it-1}$, $GPs/Hospital\ beds_{it-1}$ and $Specialists/Hospital\ beds_{it-1}$. The three concerns differ. Clearly, a crucial question is checking the extent to which results are robust to omitted variables other than those we have added to the set of regressors. To investigate this matter, we have added the full set of region specific time trends. Results are reported in Column (c), showing that, if anything, the impact of interests are larger (in absolute value) than those reported in Column (b). The endogeneity of LHA_{it-1} derives from its potential correlation with the fixed component of the error term. The bias is relevant in the "small T, large N" setup, which is not our case. The second issue relates to simultaneity and reverse causality among the variables. We have addressed that concern by lagging the variables. However, despite the previous arguments, we nevertheless estimated a full instrumental variables specification. We did so for the three regressors of interest by entering the initial values of the variables for each region as instruments for all the subsequent values, as in a GMM approach, and three predetermined regressors, namely the lagged mortality rate, the lagged infant mortality rate, and the variation in the differenced number of LHAs per capita. The results are reported in Column (d) and show that the test for exogeneity does not reject the hypothesis that the set of instruments is valid. As before, the estimated parameters are larger than those from our preferred specification.[3]

[3] We have performed a number of robustness exercises. Our results are robust to the exclusion of the non-linear combination of the other two regressors, *Specialists/Hospital beds,* from the estimating model; to the presence of supply shocks and common policy choices (such as 1999's second NHS reform); to the presence of a dummy variable for year 1995, a linear trend, a logarithmic trend and the full set of time dum-

Table 3 Results for Hypothesis 2

Model	(a)	(b)
Dependent variable	$\Delta LHA_{i,t}$	$\Delta LHA_{i,t}$
$LHA_{i,t-i}$	− 0.051***	− 0.051***
	(0.016)	(0.016)
Specialists/Hospital $beds_{i,t-1}$	0.175***	0.162***
	(0.049)	(0.048)
GPs/Hospital $beds_{i,t-1}$	− 0.298***	− 0.292***
	(0.103)	(0.102)
GPs/$Specialists_{i,t-1}$	0.083***	0.081***
	(0.029)	(0.029)
Specialists/Hospital $beds_{i,t-1} \times$ Bossi $Fini_{i,t-1}$	0.140**	
	(0.067)	
GPs/Hospital $beds_{i,t-1} \times$ Bossi $Fini_{i,t-1}$	− 0.413**	
	(0.181)	
GPs/$Specialists_{i,t-1} \times$ Bossi $Fini_{i,t-1}$	0.186**	
	(0.082)	
Bossi $Fini_{i,t-1}$	− 0.063**	
	(0.031)	
Specialists/Hospital $beds_{i,t-1} \times$ Turco $Napolitano_{i,t-1}$		0.085***
		(0.029)
GPs/Hospital $beds_{i,t-1} \times$ Turco $Napolitano_{i,t-1}$		− 0.206***
		(0.070)
GPs/$Specialists_{i,t-1} \times$ Turco $Napolitano_{i,t-1}$		0.077***
		(0.026)
Turco $Napolitano_{i,t-1}$		−0.032***
		(0.011)
F-test (31, 560) for time dummies	8.950***	8.790***
log*likelihood	−1516.50	−1517.52
Pseudo-R^2	0.1158	0.1152
Number of observations	560	560

We report results from our preferred model where the coefficients are interacted with two exogenous demand shocks, namely (a) the Bossi-Fini shock and (b) the Turco-Napolitano shock. In all the models, the set of control variables is as in Column (c) of Table 2 (parameters are not reported). Standard errors in parenthesis are robust to autocorrelation and heteroscedasticity. ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Table 3 reports the results from testing H2, namely the hypothesis that shocks to the government's perceived demand for differentiated healthcare services reinforce the incentive to consolidate in the case of complements and to keep LHAs separated in the case of substitutes. Table 3 introduces the two demand-shock proxies to our preferred specification, namely the changes in regional populations induced by the Bossi-Fini law (Column a) and the Turco-Napolitano law (Column b). Those shocks enter the set of regressors and are interacted with our relevant variables. The results show that for all of the demand shocks, the estimated coefficients of the demand proxies are statistically significant. The interaction

Footnote 3 (continued)

mies. We also have performed the empirical analysis using the within-group estimator. The results of all checks are substantially consistent with those presented here. Online Appendix 2 provides a discussion of these robustness checks and presents the associated results.

Table 4 Marginal effects

Exogenous demand shocks	(a) Absent	(b) Variable interacted with	(c)
Marginal effect on $\Delta LHA_{i,t}$ of		Bossi Fini	Turco Napolitano
$GPs/Specialists_{i,t}$	0.0302	0.0824	0.0593
	(0.012)	(0.002)	(0.000)
$GPs/Hospital\ beds_{i,t}$	− 0.1013	− 0.2306	− 0.1900
	(0.045)	(0.002)	(0.001)
$Specialists/Hospital\ beds_{i,t}$	0.0600	0.1091	0.0959
	(0.019)	(0.001)	(0.000)

Column (a) reports marginal coefficients from our preferred model. In Columns (b) and (c) we report marginal effects when the variables are interacted with the Bossi Fini law and the Turco Napolitano law variables, respectively. *P* values are shown in parenthesis

terms have the expected sign (positive for the demand for substitutes, negative for complements) and are statistically significant at the 5% level or below. These results confirm the predictions of our theoretical model. Other things being equal, an increase in the regional demand for substitute healthcare services is associated with an increase of the number of LHAs, while an increase in the demand for complementary healthcare services is associated with a decline in the number os LHAs. The incentives for consolidation and decentralization are amplified by shocks to demand.

Because the estimated Tobit coefficients represent the marginal effects of the independent variables on the latent variable (here, the desired change in the number of LHAs), we compute the marginal effects of our proxies for the demand for differentiated healthcare services on the observed outcome variable. Table 4 reports the results. As expected, the marginal effects are smaller than those derived from the Tobit estimates on the latent variable. The evidence supports H2, which predicts that the magnitudes of the estimated coefficients increase (in absolute value) when the demand shocks are considered.

5 Conclusions

This paper has analyzed the incentives of the Italian regional governments to consolidate local health authorities (LHAs). We have tested the hypothesis, derived from a stylized model, that, when a regional government perceives its citizens as demanding complementary healthcare services, it has an incentive to consolidate individual LHAs. In the case of the demand for substitute services, the regional government has an interest in dealing with a larger number of smaller LHAs. Our Tobit estimates for the years 1982–2012 show that those incentives are likely to have been in place in the Italian NHS. Over the sample period, the number of LHAs has been reduced drastically. Unquestionably, most of this reduction has been forced on the regional governments, which have the legal power to set the number of LHAs, by central government legislation and regulations since the mid-1990 s aimed at

cost-cutting. This paper has argued that, in order to gauge regional governments' incentives for consolidating, we should also look at the demand side.

The approach followed in this paper emphasizes the government's strategic choice of bureaucratic organization, namely the government's use of a control mechanism that operates ex ante (Moe 2012). However, as firstly argued by Breton and Wintrobe (1975), governments might also impose direct controls on bureaus. Such controls are costly and may take different forms of direct monitoring of bureaus associated with both different performance targets and different systems of rewards and sanctions, which, in turn, depend on regional regulations. What difference would the inclusion of control costs make to the argument presented in this paper? We would expect that the higher the cost of direct control, the more effective institutional design might be as opposed to direct control. For example, if specialists are costlier to control than general practitioners, we would expect that promoting competition between bureaus producing substitutes will be more effective than direct control.[4] Moreover, we would expect that introducing direct controls might affect our main results in so far as control costs vary between organizational forms. If controlling a consolidated agency is less costly than controlling two separate agencies (because of, say, economies of scale and scope in the control technology), consolidation might emerge as the optimal governmental choice irrespective of the nature of the healthcare services being provided. If instead government's monitoring is more efficient in smaller than in larger agencies, separation might emerge as the government's preferred choice. However, if the two potential effects of organizational form on control costs balance each other out, we would expect no change in our main findings. Following the reasoning of Coase (1937), Alchian and Demsetz (1972) and Breton and Wintrobe (1975), these different effects also might depend on the types of control device chosen. This, in turn, might also pose the issue of institutional design of bureaucratic agencies of control, or independent authorities, which we leave for future research.

Acknowledgements We would like to thank three referees for their insightful comments and suggestions, and participants to the Workshop on *Macroeconomics, rationality and institutions* held at Università La Sapienza in Rome on December 14th–15th 2017. Any mistakes are ours.

References

Alchian, A. A., & Demsetz, H. (1972). Production, information costs, and economic organization. *American Economic Review, 62,* 777–795.

Andrews, R. (2015). Vertical consolidation and financial sustainability: Evidence from English local government. *Environment and Planning C: Government and Policy, 33,* 1518–1545.

Atella, V., Belotti, F., Depalo, D., & Piano Mortari, A. (2014). Measuring spatial effects in presence of institutional constraints: The case of Italian Local Health Authority expenditure. *Regional Science and Urban Economics, 49,* 232–241.

Atella, V., & Deb, P. (2008). Are primary care physicians, public and private sector specialists substitutes or complements? Evidence from a simultaneous equations model for count data. *Journal of Health Economics, 27,* 770–785.

[4] We thank one referee for suggesting this argument.

Bagnoli, M., & McKee, M. (1991). Controlling the game: Political sponsors and bureaus. *Journal of Law Economics and Organization, 7,* 229–247.
Bates, L. J., Lafrancois, B. A., & Santerre, R. E. (2011). An empirical study of the consolidation of local public health services in Connecticut. *Public Choice, 147,* 107–121.
Bates, L. J., & Santerre, R. E. (2008). The demand for local public health services. Do unified and independent public health departments spend differently? *Medical Care, 46,* 590–596.
Blom-Hansen, J., Houlberg, K., Serritzlew, S., & Treisman, D. (2016). Jurisdiction size and local government expenditure: Assessing the effect of municipal amalgamation. *American Political Science Review, 110,* 812–831.
Bordignon, M., & Turati, G. (2009). Bailing out expectations and public health expenditure. *Journal of Health Economics, 28,* 305–321.
Breton, A., & Wintrobe, R. (1975). The equilibrium size of a budget-maximizing bureau: A note on Niskanen's theory of bureaucracy. *Journal of Political Economy, 83,* 195–207.
Breton, A., & Wintrobe, R. (1982). *The logic of bureaucratic conduct: An economic analysis of competition, exchange and efficiency in private and public organizations*. Cambridge: Cambridge University Press.
Cellini, R., Pignataro, G., & Rizzo, I. (2000). Competition and efficiency in health care: An analysis of the Italian case. *International Tax and Public Finance, 7,* 503–519.
Coase, R. H. (1937). The nature of the firm. *Economica, 4,* 386–405.
Collard-Wexler, A., Gowrisankaran, G., & Lee, R. S. (2018). 'Nash-in-Nash bargaining': A microfoundation for applied work, *Journal of Political Economy* (forthcoming).
Del Vecchio, M., & Cuccurullo, C. (2013). La dimensione ideale dell'Azienda tra economie di scala, logiche di governo e corporate identity. Quaderni FIASO, Position paper, June.
Devillanova, C. (2008). Social networks, information and health care utilization: Evidence from undocumented immigrants in Milan. *Journal of Health Economics, 27,* 265–286.
Fasani, F. (2009). Undocumented migration in Italy: A country report. CLANDESTINO project—counting the uncountable. Data and Trends across Europe–6th FP–European Commission.
Fedeli, S. (1999). Competing bureaus and politicians: A compliance approach to the diversion of public funds. *Public Choice, 100,* 253–270.
Fedeli, S. (2015). The impact of GDP on health care expenditure: The case of Italy (1982–2009). *Social Indicators Research, 122,* 347–370.
Fedeli, S., & Santoni, M. (2006). The government choice of bureaucratic organization: An application to Italian state museums. *Journal of Cultural Economics, 30,* 41–72.
Forte, F., & Powers, C. H. (1994). Applying game theory to the protection of public funds: Some introductory notes. *European Journal of Law and Economics, 1,* 193–212.
Fortney, J. C., Steffick, D. E., Burgess, J. F., Jr., Maciejewski, M. L., & Petersen, L. A. (2005). Are primary care services a substitute or complement for specialty and inpatient services? *Health Research and Educational Trust, 40,* 1423–1442.
France, G., Taroni, F., & Donatini, A. (2005). The Italian health-care system. *Health Economics, 14,* S187–S202.
Francese, M., & Romanelli, M. (2011). Health care in Italy: Expenditure determinants and regional differentials. Bank of Italy Working Paper n. 828.
Fulop, N., Protopsaltis, G., Hutchings, A., King, A., Allen, P., Normand, C., et al. (2002). Process and impact of mergers of NHS trusts: Multicentre case study and management cost analysis. *British Medical Journal, 325,* 249–252.
Garside, P. (1999). Evidence based mergers? *British Medical Journal, 318,* 445–446.
Gaynor, M., Ho, K., & Town, R. J. (2015). The industrial organization of health-care markets. *Journal of Economic Literature, 53,* 235–284.
Gaynor, M., Laudicella, M., & Propper, C. (2012). Can governments do it better? Merger mania and hospital outcomes in the English NHS. *Journal of Health Economics, 31,* 528–543.
Giannoni, M., & Hitiris, T. (2002). The regional impact of health care expenditure: The case of Italy. *Applied Economics, 34,* 1829–1836.
Goodwin, N. (2000). Leadership and the UK health service. *Health Policy, 51,* 49–60.
Horn, H., & Wolinsky, A. (1988). Bilateral monopolies and incentives to merge. *Rand Journal of Economics, 19,* 408–419.
Jacobs, R., Smith, P. C., & Street, A. (2006). *Measuring efficiency in health care. Analytic techniques and health policy*. Cambridge: Cambridge University Press.
Janssen, R. T. J. M., Leers, T., Meijdam, L. C., & Verbo, H. A. A. (2003). Bureaucracy versus market in hospital care. *Public Choice, 114,* 477–489.
Klemperer, P., & Meyer, M. (1986). Price competition vs. quantity competition: The role of uncertainty. *RAND Journal of Economics, 17,* 618–639.

Le Grand, G. (1999). Competition, cooperation, or control? Tales from the British National Health Service. *Health Affairs, 18,* 27–39.
Lester, J. P., Franke, J. L., Bowman, O'. M. A., & Kramer, K. W. (1983). Hazardous wastes, politics and public policy: A comparative state analysis. *Western Political Quarterly, 36,* 257–285.
Levaggi, R., & Zanola, R. (2003). Flypaper effect and sluggishness: Evidence from regional health expenditure in Italy. *International Tax and Public Finance, 10,* 535–547.
Maio, V., & Manzoli, L. (2002). The Italian health care system: W.H.O. ranking versus public perception. *P&T, Pharmacy and Therapeutics, 27,* 301–308.
Mapelli, V. (2012). Il sistema sanitario italiano. Bologna: il Mulino.
McDaid, D., Wiley, M., Maresso, A., & Mossialos, E. (2009). Ireland: Health system review. *Health Systems in Transition, 11,* 1–268.
Migué, J., & Bélanger, G. (1974). Toward a general theory of managerial discretion. *Public Choice, 17,* 27–43.
Miller, G. J. (1977). Bureaucratic compliance as a game on the unit square. *Public Choice, 29,* 37–51.
Moe, T. M. (1984). The new economics of organization. *American Political Science Review, 88,* 739–777.
Moe, T. M. (2012). Delegation, control, and the study of public bureaucracy. *The Forum*, 10 (2), article 4.
Nikaidô, H., & Isoda, K. (1955). Note on noncooperative convex games. *Pacific Journal of Mathematics, 5,* 807–815.
Niskanen, W. A. (1971). *Bureaucracy and representative government*. Chicago: Aldine Atherton.
Pellizzari, M. (2013). The use of welfare by migrants in Italy. *International Journal of Manpower, 34,* 155–166.
Pollitt, C. (2007). New Labor's re-disorganization. *Public Management Review, 9,* 529–543.
Schiavone, A. (2008). L'efficienza tecnica degli ospedali pubblici italiani. *Questioni di Economia e Finanza* n. 29, Banca d'Italia.
Scott, A. (1996). Primary or secondary care? What can economics contribute to evaluation at the interface? *Journal of Public Health Medicine, 18,* 19–26.
Singh, N., & Vives, X. (1984). Price and quantity competition in a differentiated duopoly. *RAND Journal of Economics, 15,* 546–554.
Talbot, C., & Johnson, C. (2007). Seasonal cycles in public management: Disaggregation and re-aggregation. *Public Money & Management, 27,* 53–60.
Ting, M. M. (2002). A theory of jurisdictional assignments in bureaucracies. *American Journal of Political Science, 46,* 364–378.
Turchetti, G. (2009). The interaction of public and private systems in healthcare provision: The Italian two-faced Janus. *European Papers on the New Welfare, 11,* 110–122.
van Dijk, C., Korevaar, J. C., Koopmans, B., de Jong, J. D., & de Bakker, D. H. (2014). The primary-secondary care interface: Does provision of more services in primary care reduce referrals to medical specialists? *Health Policy, 118,* 48–55.
Wright, D. B., & Ricketts, T. C., III. (2010). The road to efficiency? Re-examining the impact of the primary care physician workforce on health care utilization rates. *Social Science and Medicine, 70,* 2006–2010.