Centro de Economia e Finanças da UPorto

cef.up working paper
2020-02

Performance in the Delivery of Primary Health Care Services: A Longitudinal Analysis

Rita Bastião
Nuno de Sousa Pereira

Centro de Economia e Finanças da UP
www.fep.up.pt

Performance in the delivery of primary health care services: a longitudinal analysis

Rita Bastião ‡

Faculdade de Economia, Universidade do Porto

Nuno de Sousa Pereira §

Faculdade de Economia, Universidade do Porto and CEF.UP

Abstract

Primary Health Care is considered to be the cornerstone of an efficient health care system. Surprisingly, however, it has been much less studied than other levels of care. In Portugal, several reforms have been implemented to increase the efficiency and the essential role of primary health care, but their final impact is not consensual. Currently, three different organizational structures provide this level of care: Personalized Health Care Units, and two different types of Family Health Units, which were conceived as innovative settings of multidisciplinary, self-established teams with functional autonomy and a performance-based payment system. FHUs are first established as type A, but they can transit into model B and have access to additional incentives after the fulfillment of certain requirements

‡ Email address: rbastiao.phd@fep.up.pt. Financial support from FCT - Fundação para a Ciência e a Tecnologia, I.P. through the doctoral grant FRH / BD / 88354 / 2012.

§ Email address: npereira@fep.up.pt. This research has been financed by Portuguese public funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., in the framework of the project with references UIDB/04105/2020.

and after the approval by the competent authorities. Our aim is to evaluate if the units that are selected to transit to more complex organizational structures systematically exhibit better outcomes for different measures of performance, determining if the criteria used to select the winning units is appropriate. We also assess if the efficiency gaps between types of organizational structures are persistent over time.

Our dataset follows more than 800 PHC units in Portugal for the period 2009-2014. We start by conducting Simar and Wilson (2007)'s two-stage procedure applied to Data Envelopment Analysis in order to measure technical efficiency and effectiveness. We complement this analysis by considering a partial frontier approach. We also employ the dynamic concepts of window analysis and the Malmquist index decomposition to panel data in order to analyze changes in efficiency over time.

We observe that the average efficiency score of Portuguese PHC units ranges from 0.3 to 0.97, while the effectiveness score is on average around 0.9, ranging from 0.4 to 1. PHCUs significantly differ from FHUs in levels of efficiency and effectiveness, with the latter presenting better performance, particularly as FHUs-B. Units within vertical integration with hospitals are overall less efficient, but more effective in achieving specific health targets. We also observe significant geographical heterogeneity.

Keywords: Primary Health Care, efficiency, effectiveness, productivity, organizational structure

JEL Codes: I11, H21, O43, C33

1.1 Introduction

Several studies confirm the importance of Primary Health Care (PHC) for quality and equity improvements as well as for costs containment in health care systems (Atkinson and Haran, 2005; Atun, 2004; Dusheiko et al., 2008; Friedberg et al., 2010; Jürges and Pohl, 2012; Kringos et al., 2013; Macinko et al., 2003; Martin et al., 2008; Rocha et al., 2013; Starfield, 2009; Starfield et al., 2005). There is also a consensus that there is still a wide scope for efficiency gains in PHC (Amado and Santos, 2009). However, even though there has been a flourishing strand of literature on efficiency in hospital care, the same is not valid for PHC.

The Portuguese PHC has undergone important and innovative reforms in recent years. The traditional, albeit redesigned, health care centers (PHCU) and the Family Health Units (FHUs) (models A and B) are the organizational structures that provide this first level of care. PHCUs do not have financial or managerial autonomy and are managed by a Group of Primary Care Centres (ACES), under the supervision of the respective Regional Health Administration (RHAs). FHUs, on the other hand, were conceived as innovative settings of multidisciplinary and voluntarily self-established teams, with functional autonomy and a performance-based payment system (Barros et al., 2011; Pisco, 2011). FHUs include teams of three to eight general practitioners (GP), a similar number of nurses, and a variable number of administrative professionals, covering a population ranging from four to fourteen thousand individuals (Ministério-Saúde, 2013). FHUs-A usually evolve from PHCUs after the fulfillment of specific legal requirements and pre-defined standards (e.g. based on performance) and they can later on migrate to model B, having access to additional incentives. The number of new FHUs-A and B that can be established annually depends on targets set by the government[1]. Around 14% of all units (PHCUs and FHUs) belong to Local Health Units (LHUs), vertical integrated frameworks, where hospitals, primary and long-term care units are coordinated by a common management structure.

As a consequence of these reforms, the PHC service became a hybrid setting in terms of organizational structures, levels of integration and decentralization. But the outcomes of the abovementioned reforms are still dubious and scientific-based analysis of their impact is lacking (Barros et al., 2011).

These reforms can offer relevant insights to health policies in other countries. First, the Portuguese national health care system (NHS) contains a strong PHC orientation, being built around general practitioners (GPs), making Portugal one of the countries with the highest rate of GPs per capita (OECD, 2015). Second, in a comparison within 31 European countries, the Portuguese PHC is placed in the first position on structure, but at the bottom on coordination, which raises interesting questions, particularly for analyses focused on structure, process, outcomes and coordination (Kringos et al., 2013; Kringos et al., 2010a; Kringos et al., 2010b). Finally, one of the most relevant dimensions of any evaluation of PHC quality is the decline of avoidable hospital admissions for chronic diseases (Kringos et al., 2013; Macinko et al., 2010; OECD, 2015), which happened in Portugal in the follow-up of the mentioned changes.[2]

This paper extends the literature on primary health care performance by addressing the following research questions: (i) Do units that benefit from a performance-based payment systematically exhibit better performance outcomes

---

[1] Legislation by year, issued by the Ministries of Health and Finance: (Despacho-6080-B, 2014);(Despacho-4586-A, 2013);(Despacho-9999, 2012).

[2] According to the OECD (2013), the decline was from 101 avoidable admissions in chronic conditions in 2006 to 71 in 2011. More recently, Portugal is in line with EU average in this indicator (OECD, 2018).

and, if that is the case, is the conclusion uniform for different measures of performance? (ii) How persistent are efficiency gaps between groups of organizational structures over time? By answering these questions this paper contributes to a better understanding of the catalysts of an enhanced PHC system, providing important guidelines for proper policy design. To the best of our knowledge, this paper is also the first to link incentives and performance by assessing all units in a PHC system.

The empirical strategy is divided into two main parts. First, we implement Data Envelopment Analysis (DEA), a non-parametric approach, as the core methodology for performance evaluation (Charnes et al., 2013; Cooper et al., 2006; Cooper et al., 2011; Daraio and Simar, 2007; Fried et al., 2008; Ozcan, 2014; Zhu and Cook, 2007). We then incorporate Principal Component Analysis (PCA) and take into consideration Decision-Making Units' (DMUs) homogeneity (Dyson et al., 2001) to improve the discrimination power of DEA (Zhu and Cook, 2007). We overcome the drawbacks pointed out to DEA by using order-alpha partial frontier. In a second stage, we follow Simar and Wilson's approach to evaluate the determinants of efficiency across PHC units and obtain better estimation and inference (Simar and Wilson, 2007). Finally, we explore the panel structure of our data by applying Window Analysis and the Malmquist Index to study trends of efficiency gains over time across the different PHC units.

Our results yield several important findings. First, FHUs, where incentives have been employed, are consistently more efficient than PHCUs, suggesting that the new organizational settings are able to achieve the goals for which they were designed. Second, FHUs-B perform better than FHUs-A, and the differences are even stronger for FHUs-B established from the Experimental Remuneration Model

Experience[3] (RRE). We believe this can be related to a "true entrepreneurs" effect, which triggers the most incentives-oriented physicians to be the first to embrace the reforms of the new models. Third, units that belong to LHUs have lower efficiency levels, but higher effectiveness standards, which might signal coordination flaws. Fourth, when focusing on effectiveness, we observe very similar results across the considered subgroups, reaffirming the above findings. Finally, based on a dynamic analysis, efficiency and productivity remain stable over the years, which suggest that the differences across organizational types will not fade away.

Consequently, increasing the number of FHUs-B can boost overall efficiency, although this effect may be lower if a ceiling effect exists. If all PHC units reach the highest level of efficiency, it would be possible to achieve savings of close to 50% in total PHC expenditures (pharmaceuticals and diagnostic tests prescribed), based on the robust order-alpha approach.

The remainder of the paper is organized as follows. The next section provides an overview of related literature. Section 1.3 describes the data and used methods. Section 1.4 summarizes the empirical results and its discussion, and Section 1.5 presents the conclusions.

[3] The background of FHUs dates back to 1996, when the Lisbon and Vale do Tejo Regional Health Administration initiated the Alfa Project, involving an Experimental Compensation Regime (RRE) with new organizational modes and different remuneration systems in order to encourage teamwork and professional accountability (Miguel and Sá, 2010). The aim of this experiment was to decrease excessive demand in hospital emergency departments. However, in the initial stage only a limited number of professionals accessed this innovative project. The Alfa Project settled a revised general practitioner (GP) payment scheme in which groups of GPs were given overtime payments as well as other incentives in order to assure 24-hour cover and adequate referral and follow-up of patients. A preliminary internal evaluation identified enhanced patient satisfaction, a cost per patient €93 lower than a traditional health center, and a higher number of appointments (Gouveia et al., 2006).

1.2 Performance around Primary Health Care: related literature

Performance measurement is a key concern for researchers and policymakers. Most studies of performance in health care focus on hospitals and only a reduced number of them evaluates primary health care (Amado and Dyson, 2009). This may be explained by limitations in the measurement of efficiency in primary health care, such as the heterogeneity of the services performed, the failure to have easily measurable outputs, the low quality of data or the difficulty to accurately assess changes in health status of the covered patients (Hollingsworth, 2008). For example, we can have misleading results by simply considering the quantity of resources as inputs and appointments as outputs, as shorter appointments are considered as efficiency improving even though this could mean that patients were not receiving appropriate care (Amado and Dyson, 2008; Amado and Santos, 2009).

The concept of performance encompasses efficiency, effectiveness, productivity and, at a different level, equity (Amado and Dyson, 2009; Amado and Santos, 2009; Farrell, 1957; Hollingsworth et al., 1999; Mooney, 1989). Efficiency refers either to the maximization of services (output-orientation) or to the minimization of resources (input-orientation). While technical efficiency involves attaining certain goals with the least resources, allocative efficiency aims to maximize benefits from the available resources. Effectiveness, on the other hand, is related to targets and can be seen in utilitarian terms since it involves a positive contribution to the individuals' utility. Productivity is broadly defined as the ratio of output to input index. Last, equity is associated with the fair distribution of benefits across the population and it can be defined either in terms of utilization or access to health care.

Non-parametric analysis has been the main methodological approach used to measure performance in the health care sector (Hollingsworth, 2008), namely in PHC (Pelone et al., 2014). The majority of the studies are still cross sectional in nature and rely on Data Envelopment Analysis (DEA) (Amado and Dyson, 2009; Amado and Santos, 2009; Deidda et al., 2014), with only a small fraction of studies based on Stochastic Frontier Analysis (SFA) (Murillo-Zamorano and Petraglia, 2011; Olsen et al., 2013; Puig-Junoy and Ortún, 2004). Results do not significantly differ across both techniques (Giuffrida and Gravelle, 2001). It has become more common to complement DEA with other related methodologies to tackle its limitations (Cordero-Ferrera et al., 2014; Cordero et al., 2015; Ferreira et al., 2013; Giuffrida, 1999; Giuffrida and Gravelle, 2001; Staat, 2011), such as Principal Component Analysis (PCA) (Cordero-Ferrera et al., 2014; Murillo-Zamorano and Petraglia, 2011; Pääkkönen and Seppälä, 2014), partial frontiers (Cordero et al., 2015; Ferreira et al., 2013) and Malmquist indexes (Giuffrida, 1999; Staat, 2011). Previous work in PHC had a narrower scope (around 100 units), focusing on PHC units from a certain region or specific organizational setting. One of the main advantages of our empirical work is to have national coverage. The number of observations is only higher for studies on GPs.

For Portugal, Fialho et al. (2011), Ferreira et al. (2013) and Amado and Santos (2009) evaluated efficiency using simulation models and DEA. In the first two papers, only a small sample of health care units were analyzed, all from the Lisbon area. They conclude that the conversion of PHCUs into FHUs might increase efficiency in the provision of primary care services. Amado and Santos (2009) considered all health care centers, but the analysis was for a timeframe prior to the reforms that were carried out (with an efficiency score of around 0.84).

In Spain, where PHC provision shares many similitudes with Portugal, performance evaluation has been detailed more thoroughly. Most of the studies account for the influence of environmental variables on results (Cordero-Ferrera et al., 2011; Cordero et al., 2015), as well as quality concerns (Cordero et al., 2015; Murillo-Zamorano and Petraglia, 2011). Less addressed is the analysis of efficiency after the adoption of Information and Communication Technology (ICT) devices, with results showing that these are not neutral, but can improve performance of units (Deidda et al., 2014). Olsen et al. (2013) explored the relationship between organizational factors of PHC units in Denmark regarding production and efficiency, concluding that, first, the production function exhibits constant returns to scale, meaning that larger units do not lead to an increase in efficiency, and, second, that nurses are complementary inputs to GPs. For the case of England, different methodologies were addressed to measure efficiency, with results pointing out that differences across units are not relevant and that there is limited scope for productivity gains in this sector (Giuffrida, 1999; Giuffrida and Gravelle, 2001).

1.3 Methodology and Data

1.3.1 Methodological Approach

The empirical strategy involves around two complementary approaches. First, we analyze static results and compare its evolution over time using DEA, partial frontiers (order-alpha) and the impact of environmental variables on efficiency scores applying Simar and Wilson's methodology (Simar and Wilson, 2007). Second, we explore the dynamic nature of our panel dataset, to quantify changes in efficiency over time, recurring to window analysis and the Malmquist index.

1.3.1.1 Static analysis

1.3.1.1.1 Data Envelopment Analysis (DEA)

DEA is a non-parametric approach used to evaluate the performance of a set of peer entities called Decision-Making Units (DMUs) in the process of turning inputs into outputs. DEA has been extensively described in the literature (e.g. authors referred in the previous section and specialized books (Charnes et al., 2013; Cooper et al., 2006; Cooper et al., 2011; Daraio and Simar, 2007; Fried et al., 2008; Ozcan, 2014; Zhu and Cook, 2007) and applied to different industries, including the health care sector. Since the technique relies on very few assumptions and is unit free (Cooper et al., 2011), it has been used in complex environments and with multiple inputs and outputs. DEA is based on the seminal work of Charnes et al. (1978) (CCR models) and it was further developed by Banker et al. (1984) (BCC models) to account for both constant (CRS) and variable returns to scale (VRS).

The formal (and standard) representation of DEA is succinctly described (Cooper et al., 2011). Suppose we have n DMUs, where every $DMU_j, \{j = 1,2, \dots, n\}$ produces different amounts of the same s outputs, $\{y_{rj}, r = 1,2, \dots, s\}$, using the same m inputs, $\{x_{ij}, i = 1,2, \dots, m\}$. The efficiency of a specific DMU_0 can be assessed according to an optimization problem that represents a standard, input-oriented DEA model, assuming free disposability, convexity, and VRS in the so-called envelopment form (see Cooper et al. (2006)). The optimal solution of the optimization problem θ^* yields an efficiency score for a particular DMU. If $\theta^* = 1\,(< 1)$, we have an (in)efficient DMU. But the presence of boundary points is related to weak efficiency in the case of nonzero slacks. The “non-Archimedean” elements associated with $\varepsilon > 0$ handles the referred problem. It will assure that slacks are always maximized, where s_i^- and s_r^+ represent input and output slacks, respectively. Hence, the (scalar) variable θ refers to the (proportional) reduction of all inputs of the DMU being assessed (DMU_o) to increase efficiency (Charnes et al., 2013). On the other hand, λ denoting weights, that correspond to convex combinations of DMUs with output levels $Y\lambda \geq Y_0$ and input levels $X\lambda \leq X_0$. A DMU_i is fully efficient if and only if both (i) $\theta^* = 1$ and (ii) all slacks $s_i^{-*} = s_r^+ = 0$; and, (iii) λ values would be equal to 1 (Ozcan, 2014).

$$
\begin{aligned}
&\min \theta_0 - \varepsilon \left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right) \\
&s.t. \\
&\theta_0 x_{i0} = \sum_{j=1}^{n} x_{ij}\,\lambda_j + s_i^- \qquad i = 1,2,\dots,m \\
&y_{r0} = \sum_{j=1}^{n} y_{rj}\,\lambda_j - s_r^+ \qquad r = 1,2,\dots,s \\
&1 = \sum_{j=1}^{n} \lambda_j \\
&0 \leq \lambda_j, s_i^-, s_r^+ \qquad \forall i,r,j
\end{aligned}
\tag{1}
$$

The first restriction requires that the weighted sum of DMUs' inputs should equal the inputs of the DMU being evaluated. The second restriction requires that the weighted sum of the outputs should be equivalent to the focal DMU. The optimization problem becomes a CRS model in the absence of the third constraint. The relationship between the different types of returns to scale is summarized in the following table (Cooper et al., 2011):

Table1 Returns to scale

	$\theta^*_{CCR} \leq \theta^*_{NIRS} \leq \theta^*_{BCC}$	RTS	CCR Model
Case 1	If $\theta^*_{CCR} = \theta^*_{BCC}$	Constant	$\sum \lambda^*_j = 1$
Case 2	If $\theta^*_{CCR} < \theta^*_{BCC}$ then		
Case 2a	If $\theta^*_{CCR} = \theta^*_{NIRS}$	Increasing	$\sum \lambda^*_j < 1$
Case 2b	If $\theta^*_{CCR} < \theta^*_{NIRS}$	Decreasing	$\sum \lambda^*_j > 1$

Let (θ^*, λ^*) be an optimal solution for the input-oriented model in (1). Then, $(1/\theta^*, \lambda^*/\theta) = (\varphi^*, \hat{\lambda}^*)$ is optimal for the corresponding output-oriented model (Cooper et al., 2011). The formal representation of the maximization problem is the following:

$$
\begin{aligned}
&\max \varphi + \varepsilon \left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \right) \\
&s.t. \\
&x_{i0} = \sum_{j=1}^{n} x_{ij}\, \lambda_j + s_i^- \qquad i = 1,2, \dots, m \\
&\varphi y_{r0} = \sum_{j=1}^{n} y_{rj}\, \lambda_j - s_r^+ \qquad r = 1,2, \dots, s \\
&1 = \sum_{j=1}^{n} \lambda_j \\
&0 \leq \lambda_j, s_i^-, s_r^+ \qquad \forall i, r, j
\end{aligned}
\tag{2}
$$

Although new extensions and the combination of different techniques have been developed (Cooper et al., 2011), our work follows a standard specification as in Cooper et al. (2006), adopting input orientation for efficiency evaluation, and output orientation for the assessment of effectiveness.

1.3.1.1.2 Principal Component Analysis

Principal component analysis (PCA) can improve the strength of DEA models and its discrimination power by reducing the curse of dimensionality that arises in DEA in the presence of an excessive number of inputs and outputs when compared to the number of DMUs (Zhu and Cook, 2007).

Each principal component (PC) is a linear combination of the standardized values of the original variables used for the definition of an index. The number of components to be selected is conditional on the correlation of the initial variables (Cordero-Ferrera et al., 2014). Horn's parallel analysis (PA) is an empirical method for determining the number of components to be retained from PCA (Horn, 1965). The purpose is to create a random dataset with the same number of observations and variables as the original data. The rule of thumb is to select the number of components when the eigenvalues from the random data are larger than the eigenvalues from the PCA analysis.

Therefore, the DEA model in (1) is rearranged in order to use PC scores instead of the original data. The PCA-DEA formulation matches the original DEA model when the PCs explain 100% of the correlation in the original input and output matrices. The starting point is to separate $X = [X_0, X_{Lx}]$ and $Y = [Y_0, Y_{Ly}]$, where $X_0(\mathrm{Y}_0)$ represent the original values and $X_{Lx}(Y_{Ly})$ represents the linear aggregation data that become transformed through PCA. Accordingly, the formulation of (1) is now replaced by the following:

Let $L_x = \{l_{ij}^x\}$ and $L_y = \{l_{st}^y\}$ be the matrices of the PCA linear coefficients of input and output data, respectively. Now, $X_{PC} = L_x X_{Lx}$ and $Y_{PC} = L_y Y_{Ly}$ are the weighted sums of the corresponding original data, X_{Lx} and Y_{Ly}. Then:

$$
\begin{aligned}
&\min \vartheta_0 - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)\\
&\text{s. t.}\\
&\vartheta_0 x_{i0} = \sum\nolimits_{j=1}^{n} x_{ij}\,\lambda_j + s_i^-\\
&\vartheta_0 X_{PC} = X_{PC}\lambda + L_x s_{PC}^-\\
&y_{r0} = \sum\nolimits_{j=1}^{n} y_{rj}\,\lambda_j - s_r^+\\
&Y_{PC} = Y_{PC}\lambda - L_y s_{PC}^+\\
&1 = \sum_{j=1}^{n} \lambda_j\\
&0 \leq \lambda_j, s_i^-, s_r^+, s_{PC}^-, s_{PC}^+ \qquad \forall i, r, j
\end{aligned}
\tag{3}
$$

We apply this methodology in section 1.4.3.1.1.

1.3.1.1.3 Order-alpha

One of the concerns involving nonparametric envelopment estimators of frontiers is that they are very sensitive to outliers (Fried et al., 2008). To overcome this drawback, several authors have developed robust alternatives. Estimators involving the concept of a “partial” frontier avoid many of the statistical problems associated with the estimation of a traditional full frontier, while still remaining consistent estimators of the full frontier. Order-m and order-alpha integrate the approach of partial frontiers. We rely on order-alpha given the advantages that are pointed out compared to the former: first, it does not involve a re-sampling procedure (Tauchmann, 2012); second, it is continuous and not discrete, being less vulnerable to outliers (Aragon et al., 2005); and, third, it is easier to interpret since it covers the interior of the attainable set, giving a clear indication of the production efficiency (Aragon et al., 2005).

Given the focus of our work, we consider order-alpha quantile frontiers on a perspective of input orientation (for an in-depth outline see Daraio and Simar (2007); Fried et al. (2008)). For the same DMU, the benchmark will be the order-alpha quantile frontier defined as the input level not exceeded by $(1 - \alpha) \times 100$-

percent of DMUs among the overall population producing at least a level y of outputs. For α ∈ (0, 1], the α-quantile input efficiency score for the unit operating at (x, y) ∈ Ψ[4] is defined by: $\theta_\alpha(x,y) = inf\{\theta | F_{X|Y}(\theta_{X|Y}) > 1-\alpha\}$. [5]

The quantity $\theta_\alpha(x,y)$ is called the "input efficiency at level α ×100%". If α = 1, the full frontier Ψ is recovered and we would have that $\theta_1(x,y) \equiv \theta(x,y)$, which is the Debreu-Farrell input measure of efficiency (Fried et al., 2008).

Results of the implementation of this technique are discussed in section 1.4.3.1.2.

1.3.1.1.4 Outlier detection

Outliers are atypical observations that may arise from recording or measurement errors, and, therefore, should be corrected (if possible) or deleted from the data (Fried et al., 2008).

De Witte and Marques (2010) surveyed several outlier detection procedures and selected five approaches: (1) the 'leverage' concept, which addresses the disproportional influence of atypical observations (Sousa and Stošić, 2005); (2) the peer count, which consists in assessing how many observations are influenced by a certain efficient observation, with both higher and lower peer count being candidates to be outliers (Charnes et al., 1985); (3) the super-efficiency concept, that evaluates which efficient observations can simultaneously increase their inputs and undesired outputs and reduce the intended outputs by keeping themselves technically efficient (Andersen and Petersen, 1993); (4) order-m partial frontier, as developed by Simar (2003) based on the work of Cazals et al. (2002); (5) the peer

[4] Ψ is the production set.

[5] See Daraio and Simar (2007).

index, that considers the potential input reduction and output expansion as a weighted average ratio (Torgersen et al., 1996).

Observations should be screened as outliers when they prevail simultaneously as atypical by at least two of the procedures (De Witte and Marques, 2010). We rely on the first two approaches, one because it is the most recent and the other for its easy applicability. Sousa and Stošić (2005) develop the concept of leverage, measured as the standard deviation of the inefficiency estimates relative to the full sample, without considering the evaluated observation. A high value signals the presence of an influential observation. In terms of the procedure proposed by Charnes et al. (1985), the idea it to assess how many observations are influenced by a certain efficient health care unit, where the highest and lowest number are candidates, as well as the rank of repeated units.

The results yielded by the application of these procedures are discussed in section 1.4.3.1.3.

1.3.1.1.5 Second-stage procedure

DEA measures efficiency relative to a nonparametric, maximum likelihood estimate of an unobserved true frontier, based on a certain data-generating process (DGP) (Simar and Wilson, 2007). Nonparametric efficiency estimators are based on linear programming (LP) techniques for computation of estimates. They are often described as deterministic as if to suggest that the methods lack any statistical underpinnings, contrary to econometric or statistical approaches.

Performance of PHC units can be affected by external variables that are beyond their control. In order to explore the determinants of the overall efficiency, different approaches have been used. Prior studies have mostly relied on a two-stage approach, in which DEA efficiency estimates were regressed on continuous environmental variables in a parametric way, using a censored (Tobit) model (for additional references see Fried et al. (2008); Simar and Wilson (2007)). One of the major weaknesses of these approaches is the lack of a comprehensible description of a DGP, which makes inference prone to be invalid (Daraio and Simar, 2007; Dyson et al., 2001; Fried et al., 2008; Simar and Wilson, 2007). A four-stage model is one of the proposed alternatives (Cordero-Ferrera et al., 2014; Deidda et al., 2014; Fried et al., 1999), but since the dependent variables are functions of estimated efficiencies, inference problems also arise (Simar and Wilson, 2007). More importantly, DEA efficiency estimates are serially correlated.

Simar and Wilson (2007) propose a new approach to overcome the mentioned empirical challenge. The authors describe a DGP with a rational basis for regressing efficiency estimates in a second-stage analysis, becoming simultaneously possible and feasible to obtain better estimation and inference. They suggest a double bootstrap procedure, with truncated regression estimates being the correct model (Fried et al., 2008). This approach implies (i) constructing and simulating a 'sensible' DGP with (ii) artificial independent and identically distributed bootstrap samples, where (iii) standard errors and confidence intervals arise through bootstrapping/simulation (for the analytical framework see Simar and Wilson (2007)).

To our knowledge this approach has not been followed in PHC efficiency literature. Hence, this is a step forward on consistently assess the impact of external variables. Results yielded by the application of this approach are discussed in section 1.4.3.2.

1.3.1.1.6 Statistical tests based on DEA efficiency scores

This section is mainly concerned with the statistical tests that will support our framework in order to assess (i) the statistical significance between the mean efficiency score of different groups and, (ii) the validation of the relationship between inputs and outputs.

Regarding the first issue, in DEA the theoretical distribution of the efficiency scores is often unknown. Thus, we rely on nonparametric statistics that are independent of the distribution of the DEA score. Earlier literature used the rank-sum-test developed by Wilcoxon-Mann-Whitney and the Kolmogorov-Smirnov test to identify whether the differences between two groups are significant (Conover, 1999). However, since we have more than two groups, the Kruskal–Wallis test is recommended. Basically, it is a multiple-sample generalization of the two-sample Wilcoxon rank sum test to assess the hypothesis that several samples are from the same population. When the hypothesis of this test is rejected, the Dunn's test is adopted for a multiple-comparison (Dinno, 2015).

On the other hand, to validate the relationship between inputs and outputs, the Spearman's rank correlation coefficients for all pairs of variables should be considered (Conover, 1999).

1.3.1.2 Dynamic analysis

In the previous set of methodologies, each DMU was observed only once. Yet, because we have a panel dataset, we can study changes in efficiency over time.

1.3.1.2.1 Window Analysis

In Window Analysis, DEA is implemented by using a moving average analogue, which permits the identification of trends in performance (Charnes et al., 2013; Cooper et al., 2011). Accordingly, in each period, a DMU is treated as if it was a “different” unit. In other words, a DMU performance in a particular period is compared with its performance in other periods, ignoring the performance of the other DMUs (Cooper et al., 2011). Results are discussed in subsection 1.4.3.3.1.

1.3.1.2.2 Malmquist index

DEA can also be used to estimate the Malmquist productivity index (MI) (for further details see Cooper et al. (2011); Fried et al. (2008); Lee et al. (2011)). This index assesses the change in total factor productivity (TFP) on the basis of a CRS specification. TFP is the ratio of all outputs produced over all of the inputs employed to produce them. Moreover, for each observation, the geometric mean of the MI is multiplicatively decomposed into efficiency changes over time (EFFCH), commonly referred as individual catching-up, and technological progress (TECH), based on the shift of the frontier (Staat, 2011). Based on the concept of Shephard input distance function (*DI*), MI is described as the geometric mean from period *t* to period *t+1* as follows:

$$MI = EFFCH\ .TECH = \frac{D_I^{t+1}(x^{t+1},y^{t+1})}{D_I^t(x^t,y^t)} \cdot \left[\frac{D_I^t(x^{t+1},y^{t+1})}{D_I^{t+1}(x^{t+1},y^{t+1})}\frac{D_I^t(x^t,y^t)}{D_I^{t+1}(x^t,y^t)}\right]^{1/2} \quad (4)$$

In the input-oriented case, values higher (lower) than 1 imply regress (progress) in productivity, technology, or efficiency. The Malmquist decomposition goes further by allowing the breakdown of efficiency change (EFFCH) into scale efficiency (SECH) and pure efficiency change (PECH), calculated as relative to variable returns technologies, as $EFFCH = SECH\ .PECH$, which is specified in equations (5) and (6):

$$SECH = \left[\frac{D_{VRS}^{t+1}(x^{t+1},y^{t+1})/D_{CRS}^{t+1}(x^{t+1},y^{t+1})}{D_{VRS}^t(x^t,y^t)/D_{CRS}^t(x^t,y^t)} \cdot \frac{D_{VRS}^t(x^{t+1},y^{t+1})/D_{CRS}^t(x^{t+1},y^{t+1})}{D_{VRS}^t(x^t,y^t)/D_{CRS}^t(x^t,y^t)}\right]^{1/2} \quad (5)$$

$$PECH = \frac{D_{VRS}^{t+1}(x^{t+1}, y^{t+1})}{D_{CRS}^{t}(x^{t}, y^{t})} \tag{6}$$

Results are discussed in subsection 1.4.3.3.2.

1.3.2 Data

We use a panel dataset from the Central Administration of the Health System (ACSS), covering all PHC units in mainland Portugal from 2009 to 2014. The use of a complete population dataset avoids selection bias. The data comprise information on around 800 PHC units in each wave of this panel (749 in 2009; 786 in 2010; 814 in 2011; 837 in 2012; 858 in 2013; and 860 in 2014). In 2012, ACSS made available a new set of relevant variables. For this reason, static analysis is made for the last three years, while the dynamic approach considers the entire period. [6]

PHC units are organized in different structures. First, the health care units' organizational setting can be either PHCUs, FHU-A or FHU-B. Second, they are located in five regional health administrations (Norte, Lisboa, Centro, Alentejo and Algarve). Third, these units may or may not belong to LHUs (vertical integration structures)[7]. Still, PHC units provide the same basic type of services, not confounding the comparison.

The dataset characterizes the primary health care units in terms of type of organizational structure (PHCU, FHU-A or FHU-B), date of inception, date of transition to a different organizational structure, the number of doctors that practice in the unit, the number of patients with and without an assigned GP, as well as patients' ageing thresholds. Furthermore, for each unit, there is information

[6] We also considered data from 2009-2011, but results remain valid.

[7] Appendix 1 contains an overview of the organizational chart.

on utilization rate, medical appointments, health care surveillance targets (children health, family planning, maternal health, chronic conditions and risk behavior), expenditures on pharmaceuticals and diagnostic tests, as well as the share of generic drugs prescribed and referral rates[8].

Other relevant control variables were collected from the National Statistics Institute (INE). We use population density, unemployment rate, birth rate, number of hospitals, number of doctors per 1000 inhabitants, and higher education rate, all determined at the municipality level.

1.4 Results

1.4.1 Models specification

The robustness of DEA can be enhanced by comparing the results yielded by different models and variables (Amado and Dyson, 2009; Ferreira et al., 2013; Giuffrida and Gravelle, 2001; Murillo-Zamorano and Petraglia, 2011; Pelone et al., 2014). Since in primary care, the technology involved in transforming inputs into outputs is not well defined (Amado and Dyson, 2008; Pelone et al., 2014), an accurate approach needs to account for different perspectives.

Evaluation in public services should balance human resources productivity with expenditure targets. We test four alternative models (A1, B1, B2, B3). The input / output variables were chosen considering the literature review and the available data. Labor and capital have been the main input categories used in prior studies on PHC, given the growing demand for health care and the increasing expenditures under a context of constrained state budgets (Pelone et al., 2014).

[8] Patients can only visit specialists after a general practitioner referral (gate-keeping system), as defined in Mariñoso and Jelovac (2003).

Since our goal is to assess the overall efficiency of PHC units, models A1 and B1 include both human resources and expenditures as inputs. Model A1 considers human resources and number of doctors as inputs, and medical appointments and number of patients as outputs. Model B1 uses the same inputs and outputs as in A1 but adds the number of patients with vulnerable conditions as an output through a composite index. It should be noted that models identified as B consider the principal component as one of the outputs to improve the discrimination power of DEA (Dyson et al., 2001). This procedure is used in identical studies (Cordero-Ferrera et al., 2014; Murillo-Zamorano and Petraglia, 2011; Pääkkönen and Seppälä, 2014). Model B2 restricts the level of efficiency to productivity of human resources, measured by the number of doctors, and ignores costs. Finally, model B3 is concerned with cost efficiency and potential savings. Table 2 summarizes the specified models.

These approaches are consistent with the theory behind program evaluation and facilitates the analysis of the sensitivity of the results to different sets of outputs/inputs. Other argument to incorporate different models in our study is that our dataset does not fully characterize a PHC unit in terms of costs and human resources. For the former we have data on pharmaceuticals and diagnostic tests expenditures, but not personnel costs. In terms of human resources, we only have the number of doctors, but not of other professionals.

Table 2 Models specification

Description	Role	[A1] (€ | 🚹)	[B1] (€ | 🚹)	[B2] (🚹)	[B3] (€)
Expenditures					
Expendit in pharmac + diagnost tests	I	x	x		x
Doctors					
No. of doctors	I	x	x	x	
Appointments					
No. medical appointments	O	x	x	x	x
Patients					
No. patients over 65Y	O		xx	xx	xx
No. patients with GP	O	x	x	x	x
Chronic conditions					
Asthma	O		xx	xx	xx
Diabetes	O		xx	xx	xx
High blood pressure	O		xx	xx	xx
Risk behaviour					
Alcohol	O		xx	xx	xx
Obesity	O		xx	xx	xx

Notes: (xx) variables obtained through PCA; role: I – input; O – output; € - cost view; 🚹 human resources perspective.

The number of doctors and of other professionals has been largely used as input variables in prior literature analyzing efficiency in PHC (Amado and Dyson, 2008,2009; Amado and Santos, 2009; Cordero-Ferrera et al., 2014; Cordero et al., 2015; Deidda et al., 2014; Giuffrida, 1999; Giuffrida and Gravelle, 2001; Murillo-Zamorano and Petraglia, 2011). Total expenditures has also been a relevant input considered by previous authors when analyzing PHC services (Ferreira et al., 2013; Giuffrida and Gravelle, 2001; Pääkkönen and Seppälä, 2014; Staat, 2011). [9] Pharmaceuticals expenditures are based on the total amount spent by the health care unit, calculated at the retail price, and diagnosis tests are computed at the agreed price. In the literature, these inputs have been considered both separately

[9] Pharmaceuticals represent almost 80% of these expenditures.

(Amado and Santos, 2009; Cordero-Ferrera et al., 2014; Cordero et al., 2015; Deidda et al., 2014; Giuffrida, 1999; Pääkkönen and Seppälä, 2014; Staat, 2011) and jointly (Cordero-Ferrera et al., 2011; Ferreira et al., 2013; Giuffrida and Gravelle, 2001), with the former being the most common approach (Pelone et al., 2014).

We also use two sets of variables as output indicators: the number of appointments, the most common measure analyzed in PHC studies (Amado and Santos, 2009; Ferreira et al., 2013); and a set of variables that contain information about the patient. Number of patients with general practitioner has been considered to be relevant by other authors, as outlined by Amado and Dyson (2008).

Whereas the number of patients with an assigned doctor is included in all of the models, the remaining outputs are sequentially added through PCA in models B1-B3. The age structure of patients is another relevant output considered in prior studies (Amado and Dyson, 2008; Giuffrida, 1999; Pääkkönen and Seppälä, 2014; Staat, 2011), as well as chronic conditions (Amado and Dyson, 2009; Fialho et al., 2011). In parallel, we also consider risk behavior factors.

In our study, the principal components (PCs) are based on variables related to high risk patients, such as chronic conditions (diabetes and high blood pressure), risky behavior (alcohol consumption and obesity) and age distribution (Pääkkönen and Seppälä, 2014). These six outputs are included as a composite index. With the methodological approach described earlier, the construction of our synthetic index involves several steps. First, we assess the correlation between the variables. Not surprisingly, these outputs were found to be strongly correlated amongst them (Table 3). The main goal of PCA is to agglutinate the original variables into one or two single components that will be uncorrelated.

Table 3 Correlation between variables

		Asthma	Diabetes	High blood pressure	> 65Y	Alcohol	Obesity
Nr. Users with	Asthma	1					
	Diabetes	0.6907*	1				
	High blood pressure	0.7310*	0.9617*	1			
	> 65Y	0.5402*	0.8833*	0.8586*	1		
	Alcohol	0.5891*	0.4851*	0.5307*	0.3064*	1	
	Obesity	0.7802*	0.5562*	0.6172*	0.3468*	0.7153*	1

Note: correlation coefficients are significant at the 10% level or lower.

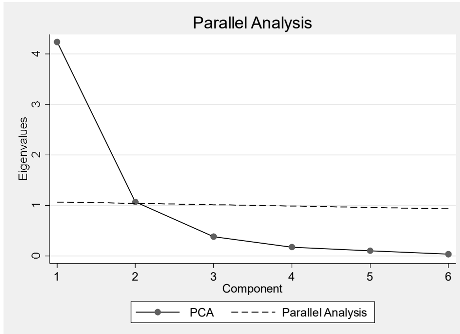
Second, as we aim to combine information from several variables into a small number of factors, we are interested in the proportion of the sample variance that is explained by the chosen PCs (Pääkkönen and Seppälä, 2014). Results reveal that, from the overall six PCs, the first two explain 88% of the total sample variance (Table 4).

Table 4 Principal component analysis

Component	Eigenvalue	Difference	Proportion	Cumulative
Component 1	4.24	3.16	0.71	0.71
Component 2	1.07	0.69	0.18	0.88
Component 3	0.38	0.21	0.06	0.95
Component 4	0.17	0.07	0.03	0.98
Component 5	0.10	0.07	0.02	0.99
Component 6	0.03	.	0.01	1.00

The parallel analysis also confirms that those two components should be retained (Figure 1). Graphical inspection shows that the dashed line crosses the solid PCA line before reaching the third component.

Figure 1 Parallel analysis



In order to interpret the PCs, it is necessary to evaluate the factor loadings. By construction, factor loadings of the first two components combined with the original variables give us two indices. These indices, weighted by the eigenvalues, end up in a composite variable used for DEA. In Table 5 it is possible to observe a strong correlation of this PC with the corresponding outputs variables.

Table 5 Spearman rank-correlation of the principal component with intrinsic outputs

		PCA	Asthma	Diabetes	l pressure	> 65Y	Alcohol	Obesity
	PCA	1						
	Asthma	0.8135*	1					
Nr. Users with	Diabetes	0.9482*	0.6907*	1				
	High blood pressure	0.9733*	0.7310*	0.9617*	1			
	> 65Y	0.8469*	0.5402*	0.8833*	0.8586*	1		
	Alcohol	0.6246*	0.5891*	0.4851*	0.5307*	0.3064*	1	
	Obesity	0.7465*	0.7802*	0.5562*	0.6172*	0.3468*	0.7153*	1

Note: correlation coefficients are significant at the 10% level or lower.

1.4.2 Descriptive statistics

Descriptive statistics are presented in Table 6, Table 7 and Table 8 for the period 2012-2014. Table 6 shows the inputs/outputs used for efficiency assessment, while Table 7 focuses on the variables for effectiveness. Columns under the title efficiency identify each variable according to its role (input or output) and wherein models they are included, while the last columns refer to statistical significance between organizational settings. Preliminary analysis suggests that health care units significantly differ from each other.[10]

Table 6 Descriptive statistics of inputs and outputs for input orientation (2012-2014)

	Overall sample			Mean by organizational setting			Efficiency		Statistical significance (t-test)		
	N	Mean	SD	PHCU	FHU-A	FHU-B	Role	Model	PHCU | FHU-A	PHCU | FHU-B	FHU-A | FHU-B
No. of doctors	2417	6.4	3.3	6.2	6.1	7.4	I1	A1, B1/B2		***	***
Expenditures in pharmaceuticals and diagnostic tests prescribed	2417	2,010,000.0	1,200,000.0	2,140,000.0	1,670,000.0	2,080,000.0	I2	A1, B1/B3	***		***
No. of overall medical appointments	2417	25,207.5	13,783.0	24,889.7	21,088.8	30,773.1	O1	A1, B1-B3	***	***	***
No. patients with GP	2417	10,865.0	5,658.5	9,854.1	10,633.0	13,611.7	O2	A1, B1-B3	***	***	***
No. patients aged more 65Y	2417	2,494.9	1,547.6	2,733.7	2,079.0	2,393.2	O3	B1-B3	***	**	***
No. patiens with asthma	2417	233.3	159.0	178.2	239.2	361.6	O4	B1-B3	***	***	***
No. registered patients with diabetes	2417	833.3	442.9	828.3	767.4	922.2	O5	B1-B3		***	***
No. registered patients with high blood pressure	2417	2,325.8	1,251.7	2,243.9	2,191.5	2,682.7	O6	B1-B3		***	***
No. patients aged ≥14Y and with excessive alcohol consumption	2417	130.4	105.2	102.0	134.3	195.4	O7	B1-B3	***	***	***
No. registered that are obese with ≥ 14 years old	2417	616.4	447.0	406.4	694.1	1,040.5	O8	B1-B3	***	***	***

Note: t-test with statistical significance between organizational types; *, **, and *** denote significant at 10%, 5% and 1% levels, respectively.

According to Table 6, PHC units have on average six doctors and eleven thousand patients with GP, and there is a significant proportion of patients above 65 years old. Expenditures on pharmaceuticals and diagnostic tests correspond to an average of 2 million euros, which emphasizes the crucial role of PHC. In terms of the different settings, FHUs of type B are generally larger, with an average of 7 doctors and around 14.000 patients with an assigned GP, as well as with a higher average number of medical appointments.

Table 7 outlines the variables used for effectiveness measurement. In terms of inputs, we observe that the different organizational settings are very similar.

[10] We also explored descriptive statistics by year, and it is observed a stable pattern.

However, there are major differences in outputs, as FHUs-B almost double or triple the outcomes obtained by PHCUs.

Table 7 Descriptive statistics of inputs and outputs for output orientation (2012-2014)

		Overall sample			Mean by organizational setting			Statistical significance (t-test)		
		N	Mean	SD	PHCU	FHU-A	FHU-B	PHCU | FHU-A	PHCU | FHU-B	FHU-A | FHU-B
Inputs	No. Patients with and without GP	2,352	12,624.59	7,365.08	13,052.43	10,829.83	13,695.78	***		***
	Enrolled women with age between 15 and 49 years old	2,352	3,215.42	2,014.92	3,309.95	2,757.51	3,522.70	***	*	***
	No. Pregnant women	2,352	37.93	24.52	30.42	37.80	55.52	***	***	***
	No. Patients that reach 1 year old	2,352	79.70	48.61	75.89	75.55	93.32	**	***	***
	No. Children with less than 2 years olds	2,352	117.57	83.30	117.79	104.30	132.30	***	***	***
	No. Patients with surveillance commitment in the diabetes program	2,352	506.90	257.47	496.40	470.48	573.16		***	***
	No. Patients with surveillance commitment in the high blood pressure program	2,352	1,927.75	1,036.75	1,841.55	1,829.69	2,240.63		***	***
Outputs	No. medical home appointments	2,352	233.21	193.05	137.90	231.68	456.22	***	***	***
	No. Women with nursing consultations in family planning	2,352	1,109.75	718.98	780.85	1,009.29	1,988.78	***	***	***
	No. Pregnant women with more than 6 nursing consultations in maternal health	2,352	30.68	22.14	21.48	30.89	51.79	***	***	***
	No. Consultations of child health of patients <330 days	2,352	419.88	248.56	350.14	410.64	592.40	***	***	***
	No. Users up to 2 years old with registration of height and weight	2,352	77.37	48.55	64.18	67.08	119.80	*	***	***
	No. Users with more than 1 nursing consultation within diabetes surveillance	2,352	389.43	225.23	318.00	394.82	549.06	***	***	***
	No. Users with record of blood pressure	2,352	1,404.80	772.95	1,106.95	1,470.65	2,020.46	***	***	***

Note: t-test with statistical significance between organizational types; *, **, and *** denote significant at 10%, 5% and 1% levels, respectively.

Table 8 exhibits the descriptive statistics for the internal and external variables used in the second-stage analysis. The dataset includes demographic (birth rate), geographic (population density), educational (higher education rate), medical (number of hospitals and doctors' density) and economic variables (unemployment rate). Internal variables comprise the units' age (and its squared), proportion of patients referred to hospitals, and two binary variables: vertical integration, which is one if the unit is within a vertical integrated structure, and list size, which is one if the doctors within the unit have, on average, less than 2500 patients allocated (Giuffrida and Gravelle, 2001; Olsen et al., 2013). This latter variable is used as a proxy for the quality of care provided.

Table 8 Descriptive statistics of internal and external variables (2012-2014)

Variables	Overall sample			Mean by organizational setting			Statistical significance (t-test)		
	N	Mean	SD	PHCU	FHU-A	FHU-B	PHCU | FHU-A	PHCU | FHU-B	FHU-A | FHU-B
Internal									
Age (years)	2,555	5.21	2.22	6.34	3.49	4.32	***	***	***
Age squared (years)	2,555	32.10	20.95	43.47	15.55	22.26	***	***	***
Dummy equal to 1 if a unit belong to LHU (vertical integration); zero otherwise	2,555	0.14	0.35	0.19	0.09	0.05	***	***	***
Dummy equal to 1 if list size is less than 2500 patients (proxy for quality); zero otherwise	2,555	0.85	0.36	0.73	0.98	1.00	***	***	**
Proportion of patients with referrals to hospitals	2,467	5.81	7.01	4.94	7.04	6.49	***	***	
External									
Population density	2,555	1,081.34	1,650.21	864.25	1,356.84	1,317.34	***	***	
Unemployment rate	2,555	9.78	2.50	9.44	9.98	10.44	***	***	***
Birth rate	2,555	7.82	1.68	7.49	8.19	8.23	***	***	
No. Hospitals	2,555	3.50	7.57	3.08	4.35	3.57	***		*
No. doctors per 1000 inhab	2,555	4.24	5.19	3.69	5.01	4.75	***	***	
Higher education rate	2,555	14.2	7.57	13.09	15.59	15.18	***	***	

Note: t-test with statistical significance between organizational types; *, **, and *** denote significant at 10%, 5% and 1% levels, respectively.

Portuguese PHC units are 5 years old on average (taking into consideration possible forms of restructuring), wherein FHUs-A are more recent (around 3 years), and PHCUs are the oldest (on average 6 years old). In terms of the variable that is used as proxy for quality, list size with less than 2500 patients, almost all FHUs fulfill this measure, against merely 73% of PHCUs. There are fewer units that belong to LHUs corresponding to 14% of the sample, but most of the ones belonging to LHUs remain as PHCUs. FHUs are more prevalent in highly densely populated areas, with more hospitals and doctors. PHCUs are commonly located in regions where higher education rates are lower.

We also included a range of indicators to characterize FHUs-B in comparison with PHCUs and FHUs-A. These indicators are defined by the health care authorities (ACSS, 2014) divided into five main sub-groups (Table 9): (i) general description (ii) efficiency, (iii) productivity, (iv) accessibility and (v) effectiveness.

Table 9 Overview of the main indicators, by type of health care unit and for the overall dataset

Topic	Variables	Overall sample			Mean by organizational setting			Statistical significance (t-test)		
		N	Mean	SD	PHCU	FHU-A	FHU-B	PHCU | FHU-A	PHCU | FHU-B	FHU-A | FHU-B
Efficiency	% users with referrals to hospitals	2467	5.81	7.01	4.94	7.04	6.49	***	***	
	% invoiced generic drugs	1668	43.48	4.92	41.48	44.39	47.32	***	***	***
	Expendit pharmaceut prescribed per user	1670	171.36	51.39	189.86	158.05	141.54	***	***	***
	Expendit diagnostic tests prescribed per user	2499	51.62	13.38	54.08	50.06	47.21	***	***	***
Productivity	% medical home appointments per 1000 users	2467	18.96	17.97	13.45	19.57	31.65	***	***	***
	% nursing home appointments per 1000 users	2467	131.23	91.51	133.11	117.50	142.69	**	***	***
Access	Rate of overall use of medical appointments	2457	66.06	11.58	64.34	64.66	71.88	**	***	***
	% utilization of medical appointments in the past	2460	81.27	12.28	80.79	76.26	88.23	***	***	***
	Overall utilization rate in the past 3Y	2403	66.36	19.11	63.01	64.10	76.83	***	***	***
Characterization	% users more 65Y	2469	20.72	6.10	22.76	19.27	17.44	***	***	***
	% users w/o GP	2469	10.83	17.70	19.37	1.81	0.57	***	***	***
	Patients per doctor	2469	2,011.28	840.67	2188.40	1776.26	1853.86	***	***	***
	% of users w/ diabetes under surveillance	2467	76.09	15.31	75.51	76.38	77.13	***	***	
	% of users w/ HBP under surveillance	2467	74.25	17.21	72.43	75.64	77.08	***	***	
	% of users w/ asthma diagnosis	2467	1.89	0.96	1.49	2.16	2.57	***	***	***
Effectiveness	% pregnant women monitored	2466	81.51	14.21	78.22	83.67	87.01	***	***	***
	% newborn monitored	2466	81.40	20.86	72.89	88.30	94.06	***	***	***
	% HBP users approp monitoring	2467	13.46	19.16	5.25	17.48	28.75	***	***	***
	% user w/ diab and heart respirat problems	2467	31.20	14.72	26.10	35.64	38.40	***	***	***
	% obese users w/ consultat past 2Y	2467	46.03	21.76	41.96	52.49	48.41	***	***	***
	% DM users approp monitoring	2467	23.27	23.00	9.87	31.61	46.16	***	***	***
	% smokers w/ consultat past 2Y	2462	24.70	21.70	20.72	27.20	31.41	***	***	***
	% pregnant w/ approp monitoring	2463	12.44	16.33	3.91	14.15	31.18	***	***	***
	% women CBA w/ approp monitoring	2467	26.77	20.85	15.60	33.61	45.96	***	***	***
	% alcohol users w/ consultat past 3Y	2466	65.58	19.62	65.63	68.14	62.48	***		***
	% children 1Y w/ approp monitoring	2466	34.34	29.79	17.50	41.38	67.08	***	***	***
	% users >25Y w/ tetanus vaccine	2467	70.67	18.81	66.08	71.90	80.41	***	***	***

Note: t-test with statistical significance between organizational types; *, **, and *** denote significant at 10%, 5% and 1% levels, respectively.

First, the percentage of users without GP is around 19% in PHCUs, while it is close to zero for FHUs. Furthermore, FHUs-B have an increased percentage of users under surveillance commitment for both diabetes and high blood pressure. In this overview, FHUs-B exhibit the lowest level of expenditures per capita in medicines and diagnostic tests prescribed, and, simultaneously, the highest ratio of generic drugs. Also, the proportion of referrals in both types of FHUs is higher than in PHCUs. As far as productivity and access is concerned, FHUs-B show the highest rates of medical/nursing home appointments and of utilization. A further point within the framework of effectiveness is the health targets achieved in FHUs, with a higher percentage of users with appropriate monitoring as compared with PHCUs, namely in the case of women in the childbearing age, pregnant women, newborn, children, patients with diabetes and high blood pressure, and patients within the risk-behavior group (alcohol consumption, smoking and obesity).

1.4.3 Comparative analysis

1.4.3.1 Cross-sectional analysis

1.4.3.1.1 Data Envelopment Analysis

We implement DEA assuming both constant returns to scale (CRS) and variable returns to scale (VRS). We then explore scale efficiency (SE) as the ratio between them (Ozcan, 2014). We compare the overall efficiency of PHC units over the years and assess efficiency gaps between the different organizational types (PHCU, FHU-A, FHU-B). For each year, all units were compared with all other units, regardless of their setting.The computed results show an average efficiency score that ranges between 0.4 and 1, depending on the specification considered. Average efficiency is only 4% higher when a VRS specification is employed, although, in the last model, the difference is much higher, revealing problems of scale inefficiency. For Portugal, previous studies reported scores between 0.8 and 0.9 (Amado and Santos, 2009; Ferreira et al., 2013). Compared with our approach, both studies use DEA, but with different data. The former is based on 2004 data, while the latter is for a small sample of 19 PHC units in the Lisbon area. In terms of inputs, Amado and Santos (2009) used the number of doctors, nurses and other staff, whereas Ferreira et al. (2013) combined working hours of professionals (doctors, nurses, administrative staff) with total costs. Outputs are very close in both studies. Moreover, Ferreira et al. (2013) enriches the analysis by considering environmental factors. In Spain, and also for PHC, identical scores to ours were found (Cordero-Ferrera et al., 2014; Cordero-Ferrera et al., 2011; Deidda et al., 2014), while, in England, higher levels of efficiency were obtained using both DEA and VRS (Giuffrida, 1999; Giuffrida and Gravelle, 2001).

Table 10 outlines the results. We offer an overview of average scores regarding the type of returns to scale, scale efficiency, as well as the share of efficient units (score of efficiency *"θ"* equals one) for each year and model. We present both CRS and VRS perspectives given the arguments for both cases in the literature. According to Cordero-Ferrera et al. (2011), assuming VRS intends to reduce potential inefficiencies due to the size of the units that are linked with scale (Charnes et al., 2013), and we know in advance that the majority of DMUs in our sample do not operate at an optimal scale (Deidda et al., 2014). However, as highlighted by Staat (2011), physicians cannot choose the number of patients they treat to become fully scale efficient. The CRS results are useful to determine the degree of scale inefficiency (Staat, 2011), which is relevant in our context. The existence of economies of scale implies that there are efficiency gains from expanding the size of units (Preyra and Pink, 2006), although there is no point of doing so in certain PHC units.

The computed results show an average efficiency score that ranges between 0.4 and 1, depending on the specification considered. Average efficiency is only 4% higher when a VRS specification is employed, although, in the last model, the difference is much higher, revealing problems of scale inefficiency. For Portugal, previous studies reported scores between 0.8 and 0.9 (Amado and Santos, 2009; Ferreira et al., 2013). Compared with our approach, both studies use DEA, but with different data. The former is based on 2004 data, while the latter is for a small sample of 19 PHC units in the Lisbon area. In terms of inputs, Amado and Santos (2009) used the number of doctors, nurses and other staff, whereas Ferreira et al.

(2013) combined working hours of professionals (doctors, nurses, administrative staff) with total costs. Outputs are very close in both studies. Moreover, Ferreira et al. (2013) enriches the analysis by considering environmental factors. In Spain, and also for PHC, identical scores to ours were found (Cordero-Ferrera et al., 2014; Cordero-Ferrera et al., 2011; Deidda et al., 2014), while, in England, higher levels of efficiency were obtained using both DEA and VRS (Giuffrida, 1999; Giuffrida and Gravelle, 2001).

Table 10 DEA results

	A1				**B1**				**B2**				**B3**			
	2012	**2013**	**2014**	**Average**	**2012**	**2013**	**2014**	**Average**	**2012**	**2013**	**2014**	**Average**	**2012**	**2013**	**2014**	**Average**
N	803	826	824		798	819	800		800	821	800		798	819	800	
CRS																
Average score	0.70	0.77	0.77	0.75	0.78	0.80	0.80	0.8	0.70	0.76	0.73	0.73	0.18	0.28	0.47	0.31
	(0.08)	(0.08)	(0.08)	(0.09)	(0.09)	(0.09)	(0.08)	(0.09)	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.09)	(0.11)	(0.15)
% efficient units	0.37%	0.61%	0.49%		2.13%	1.34%	1.25%		0.38%	0.24%	0.25%		0.25%	0.12%	0.38%	
VRS																
Average score	0.77	0.81	0.80	0.79	0.81	0.83	0.82	0.82	0.73	0.79	0.75	0.76	0.46	0.33	0.54	0.45
	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.08)	(0.09)	(0.19)	(0.14)	(0.16)	(0.19)
% efficient units	2.9%	3.1%	2.9%		4.14%	4.64%	4.00%		0.88%	1.83%	1.75%		1.63%	1.10%	1.63%	
Scale efficiency																
Average score	0.91	0.96	0.96	0.94	0.97	0.97	0.98	0.97	0.97	0.97	0.99	0.97	0.43	0.89	0.88	0.73
	(0.06)	(0.05)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.17)	(0.15)	(0.13)	(0.26)
% efficient units	0.37%	0.73%	0.49%		2.13%	1.34%	1.25%		0.38%	0.24%	0.25%		0.25%	0.12%	0.38%	
RTS (N)																
CRS	4	5	3		15	10	8		3	1	2		2	1	3	
DRS	738	653	681		587	576	674		649	691	307		775	818	511	
IRS	62	168	140		196	233	118		148	129	491		21	0	286	

Notes: Standard deviations are in parentheses for each average score; N – is the number of observations considered; % efficient units – share of units that achieve a score equal to 1. There is statistical significance between the mean efficiency score of different groups, based on Kruskal-Wallis and Dunn's test.

Our results for model B3, where the cost is the only input considered, highlight a high degree of cost inefficiency. This finding may also suggest that the main concern for PHC units is on minimizing costs rather than on increasing productivity.

As expected, when moving from a CRS to a VRS specification, the average scores are higher and the number of efficient units almost doubles. The adoption of VRS models allows us to accommodate scale effects, which might avoid bias in the case of those units that compulsory operate under a non-optimal scale of production for strategic/geographical purposes. On average, SE vary between 0.4 and 0.99. In other words, we have scale inefficiency in the interval of 0.01 up to 0.6.

Our results also show mixed evidence on the type of returns to scale. Health care managers can know which components are contributing to the inefficiency of their organization, such as the size of their operation, poor organizational factors, flow processes, or other related factors (Ozcan, 2014), as well as the local returns to scale, but in the context of PHC this is not a simple mathematical optimization problem, since equity in access to the first level of care needs to be assured. Regarding returns to scale, the majority of units were classified as exhibiting decreasing returns to scale, which draws attention to the provision of health services at a scale that is larger than the most productive one, since output levels will expand by a smaller percentage than its inputs. A DMU operating at a point where DRS holds should decrease its scale size, until the point where CRS holds (Fried et al., 2008). This result was also found in Ferreira et al. (2013), considering a small subsample of Portuguese PHC units. In contrast, for 2014 increasing returns to scale become stronger and even prevail in model B2, which might be related to administrative cleaning of patients that did not frequently used PHC units, suggesting that units might be starting to operate below the optimal scale. A DMU operating at a point where IRS holds should increase its scale size, if this is under its control, since additional input requirements may be more than compensated by a rise in output levels (Fried et al., 2008).

Furthermore, we evaluate performance by organizational setting. Results are presented in Table 11.

Table 11 DEA results, by organizational setting (VRS perspective)

(2012-2014)	Mean				Standard deviation			
	A1	B1	B2	B3	A1	B1	B2	B3
Overall sample	0.79	0.82	0.76	0.45	0.10	0.09	0.09	0.19
PHCU	0.75	0.78	0.72	0.41	0.10	0.09	0.10	0.18
FHU-A	0.82	0.84	0.77	0.43	0.07	0.07	0.06	0.16
FHU-B	0.88	0.90	0.83	0.56	0.06	0.05	0.06	0.19

Note: Results are statistically significant according to Kruskall Wallis tests.

In general, and as expected, FHUs are more efficient than the traditional health care centers in all specifications. Our results consistently show that PHCUs depict lower average efficiency scores, in contrast with the above average performance of FHUs-B, which is common to all models considered. However, since cost efficiency is one of the requirements of transitions to FHUs, the difference in model B3 is slightly higher.

It is also at the cost efficiency dimension that the difference between FHUs-B and FHUs-A is higher (0.56 versus 0.43), while the difference in the other models is only around 0.6 p.p.. One possible argument is that expenditures is one of the major goals that is monitored and established to determine additional incentives for FHUs-B, inducing a more focused behavior at this level. Additionally, PHCUs have a larger number of users above 65 years old, which hampers cost containment goals.

Our results suggest that efficiency scores are higher in the first cohort of FHUs-B, which might be explained by the fact that those who entered in the RRE had stronger intrinsic entrepreneurship capabilities. The reason for this may be related to a “true entrepreneurs” effect, leading the most incentive-oriented physicians to be the first to embrace the reforms of the new models. Those who believed and entered in the RRE were pursuing additional autonomy and incentives, given their intrinsic capabilities.

In our analysis over time, we control for unobserved characteristics of PHCUs using fixed effects, as the quality of institutions might have contributed to attract better professionals[11]. This argument is supported on the labor market literature that emphasizes matching / sorting selection. Since the seminal work of Abowd et al. (1999), positive sorting has been emphasized as a major determinant of the matching process between firms and employees. Both firms and employees have unobservable characteristics (such as quality) that may trigger a selection process such that the best workers are placed in the best firms. To explain wage variation, these authors found that employee effects are more important than firm effects. Interestingly, one other finding was that firms hiring high-wage workers are more productive, but not more profitable. In addition to this fixed effect, matching has been another stylized fact that has been recently developed (Woodcock, 2008). Unobserved heterogeneity has proven to be relevant in different labor markets. We argue that it also can be extrapolated to PHC sector, since the propensity to embrace different managerial schemes depends on intrinsic characteristics of doctors and nurses. The same way that the presence of these types of heterogeneity contributes to the bias of unexplained wage differentials, it is likely that the "best" doctors choose the most efficient units. Literature also point out that doctor's efficiency is more related with their own characteristics rather than patients' illness severity (Chilingerian, 1995). Other authors suggest that the best doctors are usually placed in larger cities, with patients preferring them due to the quality of service (Rosen, 1981).

[11] Results are presented in Appendix 2.

We conclude that FHUs-B are consistently more efficient over time[12]. Our findings also suggest that FHUs-A achieve, on average, higher efficiency scores than PHCUs. Yet, in model B3, FHUs-A do not perform better than the overall sample mean. One potential explanation is that only institutional incentives are made available to FHUs-A, which has less impact on the income of professionals than type B incentives.

When we disentangle our efficiency scores into quartiles, our results reinforce the argument of enhanced levels of efficiency within FHUs-B. In fact, in all specifications, around 60% of the FHUs-B are included in the top quartile, while 40% of the studied PHCUs are in the lowest quartile.

Table 12 shows average scores by additional groups of interest[13]. Efficiency across administrative regions are very heterogeneous. The major regional health administrations (North and Lisbon) present higher efficiency levels. For instance, for model A1, scores are 0.80 and 0.82 respectively, which compares to 0.74 in Alentejo. Arguments to sustain these findings can be explained by the fact that around half of PHC units in North and Lisbon are FHUs, given the conditions that these RHAs tend to offer. Furthermore, demographical characteristics position this two RHAs with the lowest percentage of users above 65 years, mitigating the pressure on health demand. An opposite scenario occurs in Alentejo.

Health care units that belong to LHUs perform worse. In the majority of the models, they underperform by nearly 0.05. This might signal coordination flaws,

---

[12] Our methodological approach was also conducted for each single year.

[13] For robustness, we estimate efficiency scores within homogeneous groups of units. For instance, the comparison between PHCUs in North versus PHCUs in Alentejo, or FHUs in Centro with Lisbon. For each year we separate the data into subsamples of identical units (with respect to RHA, vertical integration, seniority and size), so that we have more comparable units for benchmarking. Regardless of the used specification, this approach led to an increase of the average scores, as well as of the number of efficient units, but overall, the main conclusions regarding organizational type are similar.

which is a contradiction with the purpose of their establishment. On the other hand, as suggested by Barros and Martinez-Giralt (2003), integrated frameworks tend to trigger a higher effort on prevention to increase the overall performance of the integrated unit, which may justify lower efficiency at the primary care level. In fact, controlling for population characteristics, we observe that referrals are much lower in units that belong to LHUs. Indeed, LHUs were conceived to improve communication and coordination by integrating local hospitals with related primary care centers into a unique provider entity (Barros et al., 2011). This may perhaps explain the higher difference in model B2, as opposed to the non-existing for model B3. Controlling for population characteristics (e.g. proportion of elderly people, patients with diabetes and high blood pressure, etc.), units that belong to LHUs are 35% less likely to have referrals to hospitals[14], suggesting that the primary care component of the integrated unit absorbs costs that are born by hospitals in the absence of this integration.

Third, more recent units are likely to achieve higher efficiency scores. One possible argument in this case is the dynamic behavior of professionals seeking additional incentives. Literature on organizational inertia offers arguments that younger institutions tend to have less inertia than the older ones (Battese and Coelli, 1995; Hannan and Freeman, 1984). However, other authors argue that younger institutions can suffer from organizational instability (Colombo and Delmastro, 2002). Finally, medium and large size units, commonly located in urban areas, consistently exhibit higher efficiency levels.

[14] Standardization of variables is based on Binary Logistic Regression followed by marginal effects that contrast the rates of referrals according to the dummy variable associated with integration – Appendix 3.

Table 12 Breakdown of DEA results (VRS), by groups of interest

(2012-2014)	Average scores				Standard deviation			
	A1	B1	B2	B3	A1	B1	B2	B3
Overall sample	0.79	0.82	0.76	0.45	0.1	0.09	0.09	0.19
RHA								
Alentejo	0.74	0.79	0.70	0.45	0.12	0.10	0.09	0.17
Algarve	**0.78**	0.80	0.73	**0.50**	0.10	0.09	0.09	0.20
Centro	0.76	0.79	0.73	0.43	0.10	0.09	0.10	0.18
Lisboa	**0.82**	**0.84**	**0.78**	**0.46**	0.10	0.09	0.09	0.20
Norte	**0.80**	0.82	0.76	0.43	0.09	0.09	0.08	0.18
Vertical integration								
No	**0.80**	**0.83**	**0.77**	0.45	0.10	0.09	0.09	0.19
Yes	0.74	0.78	0.69	0.44	0.11	0.10	0.10	0.18
Seniority								
Mature	0.79	0.82	0.76	0.45	0.10	0.09	0.09	0.19
New	**0.84**	**0.87**	0.76	0.45	0.10	0.08	0.08	0.20
Size								
Small	0.74	0.79	0.71	0.34	0.11	0.10	0.10	0.14
Medium	**0.81**	**0.83**	**0.77**	0.44	0.09	0.08	0.08	0.17
Large	**0.83**	**0.85**	**0.79**	**0.60**	0.09	0.09	0.10	0.21

Note: average scores above the overall efficiency are at bold.

In the comparisons that were made, competition is not used as an argument since in Portugal patients pay user charges set by the Government and the services provided are also similar and regulated (Barros, 2017). Yet, the fact that benchmarking of units became publicly available could have played a key role.

Given the limitation of our study regarding the unavailability of labor costs, we should take into consideration that financial incentives given to FHUs may play a dual role on performance. First, it might discourage PHUCs to make an effort on performance (Tribunal-Contas, 2014), since no additional incentives are obtained. Second, FHUs-B present the highest personnel average costs given the financial incentives that they have, as opposed to their absence in PHCUs and FHUs-A (Tribunal-Contas, 2014).[15] Therefore, potential savings in terms of expenditures in

[15] These incentives might correspond to 203% of the baseline wage of health care professionals.

pharmaceuticals and diagnostic tests could be undermined by higher personnel costs.

1.4.3.1.2 Order-alpha

Table 13 summarizes the findings obtained by applying the order-alpha methodology. This approach overcomes some of the criticisms to DEA, and, therefore, provides robustness to our findings (Aragon et al., 2005). Despite having almost the same number of observations (only 0,3% were dropped for all years and models), the overall efficiency scores substantially increase when compared to the DEA results. Furthermore, we highlight the sharp rise of efficient PHC units that go from 4% to 75% of the overall sample, without impacting the observed patterns in all of the specifications. Under this procedure our findings point to a positive trend over the years, as opposed to the unstable path in DEA results.

Table 13 Efficiency evaluation according to order-alpha partial frontiers

Year	2012	2013	2014	Average
N				
A1	803	826	824	
B1	798	819	800	
B2	803	826	824	
B3	798	819	800	
Mean / (Standard deviation)				
A1	0.94	0.95	0.96	0.95
	(0.08)	(0.08)	(0.07)	(0.08)
B1	0.94	0.96	0.97	0.96
	(0.08)	(0.07)	(0.06)	(0.07)
B2	0.94	0.95	0.96	0.95
	(0.08)	(0.08)	(0.07)	(0.08)
B3	0.47	0.51	0.54	0.51
	(0.19)	(0.21)	(0.19)	(0.20)
% efficient units				
A1	56.9%	64.5%	69.1%	
B1	58.5%	64.3%	74.8%	
B2	56.9%	64.5%	69.1%	
B3	3.8%	5.7%	6.0%	

Note: Standard deviations are in parentheses for each average score; % efficient units is the share of units that achieve a score equal to 1.

Simultaneously, the results associated with differences in organizational settings are robust. FHUs are comparatively more efficient than PHCUs. As before, we explored the analysis by groups. In this case, Alentejo seems to exhibit better efficiency scores. Again, vertical integration harms efficiency, as does being a more mature provider. Table 14 presents evidence of this comparison. Figures in bold highlight the most efficient cases.

Table 14 Efficiency evaluation by groups of interest

(2012-2014)	Average scores				Standard deviation			
	A1	B1	B2	B3	A1	B1	B2	B3
Overall sample	0.95	0.96	0.95	0.51	0.08	0.07	0.08	0.20
RHA								
Alentejo	0.95	**0.97**	0.95	**0.53**	0.07	0.06	0.07	0.20
Algarve	0.95	0.94	0.95	**0.60**	0.07	0.08	0.07	0.23
Centro	0.94	0.95	0.94	0.49	0.09	0.08	0.09	0.20
Lisboa	**0.96**	0.96	**0.96**	**0.52**	0.07	0.07	0.07	0.23
Norte	0.95	0.96	0.95	0.49	0.07	0.07	0.07	0.16
Vertical integration								
No	0.96	0.96	**0.96**	0.50	0.07	0.07	0.07	0.20
Yes	0.93	0.94	0.93	**0.52**	0.09	0.08	0.09	0.20
Seniority								
Mature	0.95	0.96	0.95	0.50	0.08	0.07	0.08	0.19
New	**0.98**	**0.98**	**0.98**	**0.60**	0.06	0.06	0.06	0.22
Size								
Small	**0.98**	**0.98**	**0.98**	**0.52**	0.05	0.05	0.05	0.17
Medium	0.95	0.95	0.95	0.45	0.08	0.08	0.08	0.15
Large	0.93	0.94	0.93	**0.67**	0.08	0.08	0.08	0.26

Considering the last model, with overall expenditures of 2 billion euros, nearly half of this amount could have been saved if all PHC units were in the optimal path.

1.4.3.1.3 Outliers detection

Following the approach of De Witte and Marques (2010), and the work developed by Sousa and Stošić (2005) and Charnes et al. (1985), we combined both approaches to evaluate which observations prevail as potential outliers in each model/year. Table 15 shows the total number of observations in column 3, as considered under the context of DEA, and column 4 indicates the number of potential outliers in both procedures. The last column shows the low relative weight of these outliers, which supports the reliability of our findings[16].

Table 15 Outliers detection

Model	Year	N	Nr. Potential outliers	Relative weight of potential outliers
A1	2012	803	19	2.4%
	2013	826	16	1.9%
	2014	824	19	2.3%
B1	2012	798	23	2.9%
	2013	819	26	3.2%
	2014	800	21	2.6%
B2	2012	800	7	0.9%
	2013	821	15	1.8%
	2014	800	14	1.8%
B3	2012	798	13	1.6%
	2013	819	9	1.1%
	2014	800	13	1.6%

1.4.3.2 Second-stage approach

The data is analyzed using the estimated efficiency scores obtained with DEA and excluding observations that are signaled as potential outliers. For robustness, we perform the same analysis by subsample of organizational settings. Table 16 presents the coefficients for the overall sample, the corresponding coefficients, significance and standard error by model specification. The same approach is outlined by organizational setting.

[16] We considered DEA with and without outliers and results remain similar.

Table 16 Second-stage analysis

			A1		B1		B2		B3	
			Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Overall models	Internal variables	Age (years)	0.20	(0.45)	0.71	(0.41) *	1.70	(0.39) ***	1.65	(0.82) **
		Age (years) squared	-0.12	(0.05) ***	-0.19	(0.04) ***	-0.26	(0.04) ***	-0.17	(0.09) *
		Dummy equal to 1 if a unit belong to LHU (vertical integration); zero otherwise	-4.28	(0.60) ***	-2.97	(0.54) ***	-5.38	(0.52) ***	4.45	(1.08) ***
		Dummy equal to 1 if list size is less than 2500 patients (proxy for quality); zero otherwise	1.53	(0.68) **	-0.35	(0.60)	-0.93	(0.57)	4.91	(1.22) ***
		Proportion of patients with referrals to hospitals	0.21	(0.03) ***	0.18	(0.03) ***	0.03	(0.02)	0.91	(0.05) ***
	External variables	Population density	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00) *
		Unemployment rate	0.58	(0.09) ***	0.45	(0.08) ***	0.47	(0.08) ***	-0.31	(0.15) **
		Birth rate	0.96	(0.17) ***	0.66	(0.16) ***	0.45	(0.16) ***	2.65	(0.32) ***
		No. Hospitals	-0.04	(0.05)	-0.03	(0.05)	-0.03	(0.04)	-0.07	(0.10)
		No. doctors per 1000 inhab	-0.30	(0.08) ***	-0.22	(0.08) ***	-0.24	(0.07) ***	0.06	(0.15)
		Higher education rate	0.20	(0.06) ***	0.09	(0.06)	0.15	(0.05) ***	-0.23	(0.11) **
PHCU	Internal variables	Age (years)	-2.28	(0.75) ***	-1.33	(0.65) **	-0.59	(0.70)	-3.73	(1.46) **
		Age (years) squared	0.28	(0.08) ***	0.13	(0.07) *	0.11	(0.07)	0.43	(0.15) ***
		Dummy equal to 1 if a unit belong to LHU (vertical integration); zero otherwise	-3.31	(0.67) ***	-1.84	(0.59) ***	-4.90	(0.61) ***	4.85	(1.25) ***
		Dummy equal to 1 if list size is less than 2500 patients (proxy for quality); zero otherwise	-3.10	(0.67) ***	-4.57	(0.61) ***	-4.44	(0.63) ***	-0.10	(1.21)
		Proportion of patients with referrals to hospitals	-0.01	(0.05)	-0.01	(0.05)	-0.09	(0.05) *	0.65	(0.10) ***
	External variables	Population density	0.00	(0.00) ***	0.00	(0.00) **	0.00	(0.00) ***	0.00	(0.00)
		Unemployment rate	0.35	(0.10) ***	0.29	(0.09) ***	0.34	(0.10) ***	-0.63	(0.19) ***
		Birth rate	0.31	(0.20)	0.12	(0.18)	-0.01	(0.20)	1.76	(0.39) ***
		No. Hospitals	0.10	(0.07)	0.11	(0.06) *	0.07	(0.07)	0.19	(0.14)
		No. doctors per 1000 inhab	-0.37	(0.11) ***	-0.21	(0.10) **	-0.36	(0.11) ***	0.20	(0.21)
		Higher education rate	0.25	(0.08) ***	0.05	(0.08)	0.33	(0.08) ***	-0.43	(0.16) ***
FHU-A	Internal variables	Age (years)	-2.73	(0.61) ***	-2.92	(0.59) ***	0.86	(0.50) *	0.75	(1.17)
		Age (years) squared	0.29	(0.07) ***	0.31	(0.07) ***	-0.10	(0.06) *	0.00	(0.14)
		Dummy equal to 1 if a unit belong to LHU (vertical integration); zero otherwise	-0.32	(0.93)	-0.89	(0.94)	-1.78	(0.81) **	8.14	(1.91) ***
		Dummy equal to 1 if list size is less than 2500 patients (proxy for quality); zero otherwise	0.08	(3.89)	-0.40	(4.70)	-2.83	(4.17)	15.66	(7.76) **
		Proportion of patients with referrals to hospitals	0.10	(0.03) ***	0.09	(0.03) ***	-0.06	(0.03) **	0.83	(0.07) ***
	External variables	Population density	0.00	(0.00)	0.00	(0.00) **	0.00	(0.00) *	0.00	(0.00)
		Unemployment rate	0.23	(0.12) **	0.17	(0.11)	0.13	(0.10)	0.02	(0.24)
		Birth rate	0.91	(0.26) ***	0.43	(0.26)	-0.17	(0.22)	3.00	(0.55) ***
		No. Hospitals	0.05	(0.07)	0.05	(0.07)	0.11	(0.06) **	-0.08	(0.13)
		No. doctors per 1000 inhab	-0.38	(0.10) ***	-0.33	(0.10) ***	-0.39	(0.09) ***	-0.11	(0.22)
		Higher education rate	0.15	(0.07) **	0.11	(0.07)	0.06	(0.07)	-0.02	(0.15)
FHU-B	Internal variables	Age (years)	0.76	(0.60)	1.24	(0.54) **	1.09	(0.54) **	1.55	(1.95)
		Age (years) squared	-0.05	(0.07)	-0.10	(0.07)	-0.11	(0.07)	-0.02	(0.24)
		Dummy equal to 1 if a unit belong to LHU (vertical integration); zero otherwise	-5.27	(1.00) ***	-3.20	(0.96) ***	-4.11	(0.99) ***	2.55	(3.35)
		Dummy equal to 1 if list size is less than 2500 patients (proxy for quality); zero otherwise							-0.59	(51.58)
		Proportion of patients with referrals to hospitals	0.01	(0.03)	0.00	(0.03)	-0.23	(0.03) ***	0.93	(0.11) ***
	External variables	Population density	0.00	(0.00) **	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
		Unemployment rate	0.32	(0.11) ***	0.14	(0.11)	0.22	(0.10) **	-0.87	(0.39) **
		Birth rate	0.56	(0.30) *	0.95	(0.26) ***	-0.12	(0.27)	4.40	(1.01) ***
		No. Hospitals	0.09	(0.06)	0.04	(0.06)	0.03	(0.06)	0.02	(0.21)
		No. doctors per 1000 inhab	-0.22	(0.10) **	-0.17	(0.10) *	-0.12	(0.09)	-0.05	(0.33)
		Higher education rate	0.00	(0.07)	-0.07	(0.07)	-0.08	(0.07)	-0.35	(0.24)

Notes: *, **, and *** denote significant at 10%, 5% and 1% levels, respectively; the set of results is shown as multiplied by 100.

The age of health care units is more likely to positively affect performance, in the overall sample, which indicates that more mature institutions tend to be more efficiency-driven. However, in PHCUs and FHUs-A, mixed evidence is observed. Yet, the negative effect of age squared means that as health care units get older the effect of age is lessoned. Studies on different fields also point out mixed results, arguing organizational inertia in older institutions, but learning experience to overcome obstacles as opposed to less consolidated structures in younger institutions (Battese and Coelli, 1995; Colombo and Delmastro, 2002; Hannan and Freeman, 1984).

In terms of the coefficient associated with vertical integration, it is usually negative in the models related to human resources, while cost efficiency (B3) is benefited in units that belong to a LHU, which signals the concern with costs. In fact, it is argued that investment in preventive care increases when shifting from an autonomous to an integrated management (Barros and Martinez-Giralt, 2003). A related variable in this setting is the proportion of referrals in a gatekeeping system. Our results reveal a dual effect, but more commonly a high percentage of referrals tend to favor efficiency. Further analysis would be interesting in order to foil dubious referrals emergencies. In general this variable has been used as input by other authors (Amado and Dyson, 2008; Olsen et al., 2013; Staat, 2011).

The last internal variable considered, as a proxy for quality, is the number of patients per doctor to be less than 2,500 users. As emphasized by Giuffrida (1999) and Olsen et al. (2013), that used this variable in assessing efficiency of PHC in UK and Denmark, respectively, it is assumed that GPs with shorter lists are able to deliver improved services and easier access for their patients. Our findings confirm the benefit for efficiency of this variable across models. But, for the subsamples of FHUs-B, the coefficient was not statistically significant. In contrast, in the sample of FHUs-A, the estimated effect is negative in terms of the model that is associated with human resources (B2) and positive in terms of cost efficiency (B3), which is puzzling.

In terms of external variables, we considered population density to account for geographical features (Cordero-Ferrera et al., 2014; Cordero-Ferrera et al., 2011; Ferreira et al., 2013; Puig-Junoy and Ortún, 2004). Our findings suggest little influence of this variable in efficiency (Ferreira et al., 2013; Puig-Junoy and Ortún, 2004), with trend towards a negative sign. However, in model B3, the coefficient is positive and statistically significant, suggesting that regions with a higher density, usually large cities, have more efficient units (Cordero-Ferrera et al., 2014; Cordero-Ferrera et al., 2011), which can be associated with the coexistence of more private alternatives. Given the correlation of population density with urban areas and gross domestic product we do not include them. Birth rate positively affects efficiency in a consistent way, although previous studies only obtained a weak impact (Cordero-Ferrera et al., 2014; Cordero-Ferrera et al., 2011).

Our findings suggest that unemployment rate positively affects the estimated efficiency scores in all models. This outcome is transversal to different models and subsamples. One of the arguments is the postponement of preventive appointments given financial constraints. This artificially increases efficiency scores in the short run, but with possible serious implications in a near future. Prior literature did not find this variable to be significant (Staat, 2011).

In terms of health indicators, we consider the number of hospitals, as well as the number of doctors per 1000 inhabitants. Despite a weak influence of these variables on efficiency levels, the number of professionals negatively influences performance, while the opposite is true for the case of hospitals. A potential reasoning is that municipalities with a high ratio of professionals have hospitals or other alternatives, which deviates workforce from PHC units. Conversely, by its proximity, this type of health care units concentrates more serious illnesses. In the context of Portugal, Ferreira et al. (2013) found a positive relationship between the distance to the reference hospital arguing that, when a hospital is located farther away from population, more patients will go to PHC.

The relationship between education and health has a long history in health economics (see Grossman (1972)), although previous studies did not consider it for efficiency purposes. Our findings suggest mixed evidence, although with weak influence.

1.4.3.3 Dynamic analysis

1.4.3.3.1 Window analysis

Table 17 outlines efficiency estimates considering a three-year window under VRS. The main goal is to assume that all units are different in each year and compare them to the overall set of DMUs.

Each DMU is represented as if it was a different DMU for each of the successive years in the first window, consisting of the year at the top of the following table. The window is then shifted one year and the process continues in this manner, shifting the window forward one period each time and concluding with the final analysis. Table 17 illustrates the results of this analysis in the form of efficiency scores, where the structure of this table depicts the framework of the analysis. For instance, in the third "window" (2009-2011) it is represented the constraints of the DEA model as though it was a different DMU in years 2009, 2010 and 2011.

The way results are displayed enables the identification of trends in performance, as well as the stability of reference sets. "Row views" clarify performance trends.

Table 17 Window analysis

Window	2009	2010	2011	2012	2013	2014	Average
Model 1							
2009 - 2009	0.90						
2009 - 2010	0.80	0.78					
2009 - 2011	0.79	0.77	0.76				
2010 - 2012		0.76	0.74	0.75			
2011 - 2013			0.75	0.75	0.76		
2012 - 2014				0.75	0.76	0.77	
Window analysis effic average	0.83	0.77	0.75	0.75	0.76	0.77	0.76
Model 8							
2012 - 2012				0.81			
2012 - 2013				0.79	0.81		
2012 - 2014				0.77	0.78	0.81	
Window analysis effic average				0.79	0.80	0.81	0.79
Model 9							
2012 - 2012				0.73			
2012 - 2013				0.73	0.73		
2012 - 2014				0.71	0.71	0.73	
Window analysis effic average				0.71	0.71	0.73	0.72
Model 10							
2012 - 2012				0.46			
2012 - 2013				0.29	0.31		
2012 - 2014				0.28	0.31	0.32	
Window analysis effic average				0.34	0.31	0.32	0.33

Note: windows analysis efficiency average refers to the results reported by three-year window.

Briefly, the column view allows comparison of DMUs across different reference sets and therefore gives additional insights on the stability of these scores as the reference sets change. In our case, we observe a stability pattern. On the other hand, the row view reflecting performance trends suggests a steady behavior. Overall, it is possible to observe stability and a steady behavior of efficiency scores over the years.

1.4.3.3.2 Malmquist index

Based on the contribution of DEA, results for the Malmquist index are depicted in Table 18. Our results suggest that productivity remains by and large constant, supporting findings by other authors (Staat, 2011).

The first column identifies the period of analysis, and the second refers to the model considered in DEA. The remaining columns decompose the total factor productivity into efficiency change and catching-up effect (multiplicative effect of the changes in pure and scale efficiency). It should be noted that we restrict our sample to DMUs that remain under the same organizational setting in these years (around 80% of the overall sample).

Overall, we observe progress in total factor productivity in both periods. While in 2012/2013 the leading cause is efficiency change with a positive catching-up effect, in 2013/2014 the breakdown of factor productivity is balanced between efficiency change and technological progress.

Furthermore, our results show technological changes in both periods, with roll-back in 2012/2013 and progress in 2013/2014. Innovation can be associated with better procedures and/or IT devices brought to clinical practices to allow monitoring and electronic prescriptions. In practical terms, a regulatory change occurred in June 2012, when it began to be mandatory the prescription under the international non-proprietary name (INN).

These conclusions should not be considered lightly as a 3-year window cannot be taken as a long-term trend, and changes in patient health would have been a better output measure if available.

We also obtain identical findings by organizational setting. Thus, while FHUs attain a slight improvement in productivity progress when compared to PHCUs, technological change seems to act as disadvantage to FHUs performance improvement.

Table 18 Malmquist index decomposition

Period	Model	Total factor productivity	Technical efficiency change (CRS)	Technological change	Pure efficiency change (VRS)	Scale efficiency change (SE)
2012-2013	A1	0.98	0.94	1.05	0.96	0.98
	B1	0.98	0.98	1.00	0.98	1.00
	B2	0.99	0.96	1.03	0.95	1.00
	B3	0.89	0.38	2.34	0.81	0.47
	Average	0.85	0.72	1.19	0.92	0.78
2013-2014	A1	0.98	0.99	0.99	1.01	0.99
	B1	0.95	1.00	0.95	1.01	0.99
	B2	0.98	1.04	0.94	1.06	0.98
	B3	0.91	0.87	1.05	0.91	0.95
	Average	0.95	0.97	0.98	0.99	0.98
Overall period	**Average**	**0.89**	**0.83**	**1.08**	**0.96**	**0.87**
	Model					
	A1	0.98	0.96	1.02	0.98	0.98
	B1	0.96	0.99	0.98	1.00	0.99
	B2	0.98	1.00	0.99	1.00	0.99
	B3	0.78	0.54	1.45	0.86	0.63
	Organizational type					
	PHCU	0.88	0.82	1.07	0.95	0.87
	FHU-A	0.90	0.84	1.07	0.95	0.88
	FHU-B	0.90	0.83	1.08	0.98	0.85

Note: group differences are statistically significant only in models A1 and B2, in the case of FHUs.

1.4.4 Effectiveness

We also assess performance through the analysis of effectiveness. We apply an output orientation since the main goal is to assess the increase of outputs' level while keeping the inputs' level unchanged. Hence, according to Mooney (1989), effectiveness is more perceived from the patient's point of view, while efficiency and equity are more linked to an organizational improvement. Moreover, it is a challenge to address outcomes and quality rather than efficiency measures (Ozcan, 2014).

For this purpose we consider seven inputs and the same number of outputs, each one linked to specific health targets, as defined by ACSS (Table 19).

Table 19 Effectiveness assessment of the inputs and outputs

Inputs	**Outputs**	**Health target**
No. Patients with and without GP	No. medical home consultations	Home consultations
Enrolled women with age between 15 and 49 years old	No. Women with nursing consultations in family planning	Family planning
No. Pregnant women	No. Pregnant women with more than 6 nursing consultations in maternal health	Maternal health
No. Patients that reach 1 year old	No. Consultations of child health of patients <330 days	Children health
No. Children with less than 2 years olds	No. Users up to 2 years old with registration of height and weight	Children health
No. Patients with surveillance commitment in the diabetes program	No. Users with more than 1 nursing consultation within diabetes surveillance	Diabetes health target
No. Patients with surveillance commitment in the high blood pressure program	No. Users with record of blood pressure	High blood pressure

Results are presented in Table 20 for original variables and Table 21 for PC, based on DEA. Given the reduction of the number of observations in the first case, we could infer that the curse of dimensionality may be affecting the results. As suggested by Pääkkönen and Seppälä (2014), PCA limits the influence of noise and avoid multicollinearity. In Table 22 results are estimated according to order-alpha partial frontier. For robustness, we consider both the original variables and PC (inputs and outputs), following the steps described earlier. Our results confirm FHUs' higher performance when compared to PHCUs. Furthermore, FHUs-B are more effective in achieving the health targets defined in the contract. From the perspective of effectiveness, a balance sample between IRS and DRS is observed, as opposed to what was observed in efficiency evaluation.

Table 20 DEA results, according to original variables (2012-2014)

	2012-2014			2012			2013			2014		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Overall sample												
VRS	1286	0.87	0.11	416	0.86	0.12	446	0.87	0.11	424	0.88	0.10
CRS	1196	0.85	0.12	356	0.84	0.13	395	0.85	0.12	445	0.87	0.11
Scale	1053	0.98	0.05	314	0.98	0.05	355	0.98	0.05	384	0.98	0.05
Type of returns												
CRS	115			26			36			53		
DRS	569			155			219			195		
IRS	602			235			191			176		
Organizational type												
VRS												
PHCU	805	0.84	0.12	274	0.82	0.12	277	0.84	0.12	254	0.85	0.11
FHU-A	375	0.91	0.08	115	0.92	0.07	127	0.91	0.08	133	0.91	0.08
FHU-B	106	0.97	0.02	27	0.97	0.02	42	0.97	0.03	37	0.97	0.02
CRS												
PHCU	734	0.81	0.12	240	0.79	0.13	240	0.80	0.12	254	0.83	0.12
FHU-A	328	0.90	0.08	89	0.90	0.08	102	0.89	0.09	137	0.91	0.08
FHU-B	134	0.97	0.02	27	0.97	0.02	53	0.97	0.02	54	0.97	0.02
Scale												
PHCU	664	0.97	0.06	213	0.97	0.06	222	0.97	0.06	229	0.97	0.06
FHU-A	298	0.99	0.01	82	1.00	0.01	96	0.99	0.01	120	0.99	0.02
FHU-B	91	1.00	0.01	19	1.00	0.00	37	1.00	0.00	35	1.00	0.01

Note: N – number of observations; SD – standard deviation.

Table 21 DEA results, according to PCA (2012-2014)

	2012-2014			2012			2013			2014		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Overall sample												
VRS	2,352	0.54	0.20	756	0.52	0.21	805	0.56	0.2	791	0.54	0.19
CRS	2,352	0.38	0.14	756	0.38	0.15	805	0.39	0.14	791	0.38	0.13
Scale	2,352	0.72	0.11	756	0.74	0.12	805	0.71	0.1	791	0.71	0.10
Type of returns												
CRS	3			1			1			1		
DRS	2,299			747			783			769		
IRS	50			8			21			21		
Organizational type												
VRS												
PHCU	1,221	0.42	0.16	405	0.40	0.16	420	0.45	0.16	396	0.43	0.16
FHU-A	605	0.58	0.14	189	0.57	0.14	204	0.59	0.14	212	0.57	0.13
FHU-B	526	0.77	0.11	162	0.78	0.10	181	0.78	0.09	183	0.74	0.12
CRS												
PHCU	1,221	0.31	0.13	405	0.29	0.13	420	0.32	0.13	396	0.30	0.12
FHU-A	605	0.42	0.10	189	0.42	0.10	204	0.42	0.10	212	0.40	0.09
FHU-B	526	0.53	0.07	162	0.55	0.06	181	0.53	0.06	183	0.50	0.08
Scale												
PHCU	1,221	0.73	0.14	405	0.75	0.15	420	0.72	0.13	396	0.72	0.13
FHU-A	605	0.73	0.05	189	0.76	0.06	204	0.71	0.05	212	0.71	0.04
FHU-B	526	0.69	0.04	162	0.71	0.04	181	0.68	0.03	183	0.68	0.03

Note: N – number of observations; SD – standard deviation.

Table 22 Results based on order-alpha (2012-2014)

Order-alpha	N	Mean	SD	Min	Max
Overall sample	2,341	0.99	0.04	0.53	1.00
Y					
2012	752	0.99	0.05	0.54	1.00
2013	802	0.99	0.05	0.53	1.00
2014	787	0.99	0.04	0.56	1.00
Organizational type					
PHCU	1,220	0.98	0.06	0.53	1.00
FHU-A	595	1.00	0.02	0.74	1.00
FHU-B	526	1.00	0.00	1.00	1.00

Note: N – number of observations; SD – standard deviation.

1.5 Concluding remarks

The present study assesses the determinants of performance in the Portuguese PHC units and explores the impact of the adopted organizational setting on performance dynamics.

The findings of our research are quite robust to different subsamples and comparison groups, as well as methodological approaches. The following conclusions can be drawn: (i) FHUs are simultaneously more efficient and effective than PHCUs; (ii) FHUs-B are the organizational type with the best overall performance; (iii) units belonging to Local Health Units consistently reveal lower efficiency scores, which may suggest a stronger focus on preventive care and problems of coordination.

Regarding methodological concerns, order-alpha results should be considered as the upper threshold por policy evaluation, given its robustness. Therefore, if all PHC units reach the highest level of efficiency, it would be possible to achieve potential savings of close to 50% in total PHC expenditures, which is economically meaningful. Nevertheless, DEA remains widely used (yielding identical scores), with additional advantages to pursue complementary approaches.

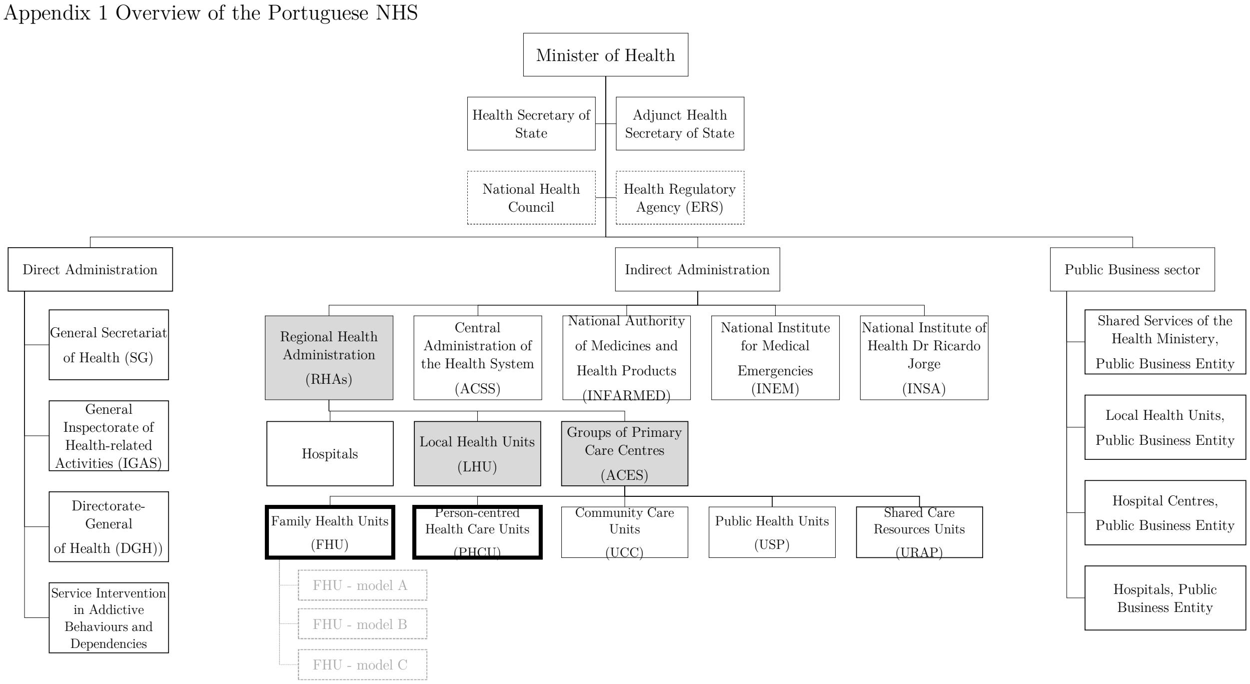
Overall, under a dynamic perspective, our results seem to indicate that productivity remains by and large constant. Furthermore, window analysis allows us to observe a steady behavior of efficiency over the years.

From the second-stage approach, our findings also suggest that vertical integration in the form of LHUs are likely to negatively influence efficiency, while favorably affecting cost efficiency. In terms of institutions' age, it seems that experience has a positive impact on efficiency by providing a more stable framework until a certain age. On the other hand, the other internal variables, such as patients list size and percentage of referrals reveal mixed influence depending on the model considered and the organizational subset. In terms of external variables, population density reveals a small and mainly negative impact on efficiency, which might indicate higher pressure from health care demand. The unemployment rate has a positive impact on efficiency which may indicate postponement of treatment. Finally, regarding health care supply determinants, the number of hospitals by municipality improves PHC units performance, while the number of doctors per 1000 inhabitants has an opposite effect.

Finally, our findings lend strong support to the implementation of FHUs, in particular type B. The on-going reforms should continue to foster the implementation of these organizational structures, which will result in potential savings and higher levels of effectiveness.

The current investigation was mainly limited by the absence of some variables such as the number of nurses and personnel expenditures. In fact, the impact of the latter is one of the current criticisms on evaluating potential savings of FHUs. Yet, our results are robust in terms of obtaining improved levels of efficiency and effectiveness as far as human resources is concerned. As a final remark, further work needs to be developed to establish whether the existence of a ceiling effect constrains future performance improvements.

Appendix 1 Overview of the Portuguese NHS



Source: Own elaboration. Based on Barros et al. (2011) and website of the Portuguese Health Ministry (accessed in 2015).
Notes: grey boxes correspond to the core structures.

Appendix 2 Fixed effects estimation (2012-2014)

Model 1	Coef	Bootstrap Std. Error	z	P> | z |	[95% Conf. Interval]	
Quality proxy (<2500 patients)	-0.02	0.01	-1.80	0.07	-0.04	0.00
Population density	0.00	0.00	-4.52	0.00	0.00	0.00
Y						
2013	0.02	0.00	6.65	0.00	0.01	0.02
2014	0.01	0.00	3.64	0.00	0.01	0.02
PHCU interacted with year						
2013	0.02	0.01	3.03	0.00	0.01	0.03
2014	0.01	0.01	2.43	0.02	0.00	0.02
FHU-A interacted with year						
2013	0.02	0.01	2.17	0.03	0.00	0.03
2014	0.01	0.01	1.51	0.13	0.00	0.03
Rho	0.93					

Notes: i) Equation estimated for the score of efficiency, controlling for fixed effects of each unit, including the type of unit varying with time. Thus, interacting PHCUs with year dummies, and FHUs-A with year dummies, etc. Each coefficient would give the efficiency differential between units, comparing it with the one that stay out (FHU-B), in each period. Results remain, so it could be argued that the fixed effect includes the quality of the institution, not observed, but that probably has attracted better doctors. It should be noted that one of the basic assumptions is that this "quality" of the unit is assumed to be fixed / permanent / time-invariant over time in fixed-effect models.

Appendix 3 Referrals controlling for population characteristics (2012-2014)

	N	Mean	Std	Min	Max
Outside LHU	2,113	6.07	7.17	0.00	42.12
Within LHU	354	4.26	5.71	0.00	23.90

The sample average of referrals of units that belong to LHUs is 30% lower

Margins after Generalized Linear Models, controlling for risk and age:

	df	chi2	P>chi2	
Dummy LHU	1	10.86	0.001	
	Contrast	S.E.	[95% Conf.	Interval]
Within vs outside LHU	-1.35	0.41	-2.15	-0.55

Controlling for population characteristics, referrals from units that belong to LHUs tend to be, on average, 35% lower than PHC units that only belong to ACES. Hence, the contrast estimate of -1.35 indicates that unconditional on group, units of LHUs on average are about 35% less likely than the others to have referrals to hospitals. Moreover, the chi-squared statistic shows that the contrast is significantly different from zero.

Appendix 4 Efficiency scores based on different subsamples (VRS, 2012-2014)

		Model A1			Model B1		Model B2		Model B3		
		N	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**N**	**Mean**	**S.D.**
Subsample based on RHA	Overall scores	2453	0.85	0.10	2417	0.87	2421	0.83	2417	0.58	0.20
	PHCU	1301	0.82	0.11	1284	0.84	1287	0.80	1284	0.54	0.20
	Alentejo	126	0.82	0.13	120	0.87	120	0.83	120	0.70	0.18
	Algarve	51	0.93	0.10	51	0.94	51	0.90	51	0.80	0.22
	Centro	311	0.86	0.09	308	0.88	309	0.81	308	0.57	0.22
	Lisboa	360	0.83	0.09	358	0.84	360	0.79	358	0.48	0.18
	Norte	453	0.77	0.09	447	0.81	447	0.78	447	0.49	0.14
	FHU-A	619	0.86	0.07	609	0.88	610	0.83	609	0.56	0.18
	Alentejo	32	0.91	0.09	32	0.93	32	0.90	32	0.80	0.19
	Algarve	18	0.89	0.05	17	0.92	17	0.85	17	0.72	0.21
	Centro	88	0.88	0.05	87	0.89	87	0.84	87	0.51	0.20
	Lisboa	200	0.89	0.07	192	0.90	193	0.83	192	0.55	0.21
	Norte	281	0.83	0.06	281	0.85	281	0.82	281	0.55	0.13
	FHU-B	533	0.91	0.06	524	0.93	524	0.89	524	0.67	0.17
	Alentejo	12	0.96	0.05	12	0.96	12	0.95	12	0.82	0.15
	Algarve	9	0.99	0.02	8	0.99	8	0.97	8	0.88	0.16
	Centro	45	0.96	0.04	43	0.96	43	0.91	43	0.62	0.15
	Lisboa	156	0.92	0.05	150	0.94	150	0.86	150	0.64	0.21
	Norte	311	0.90	0.06	311	0.92	311	0.89	311	0.69	0.14
Subsample based on seniority	Overall scores	2453	0.80	0.11	2417	0.83	2421	0.77	2417	0.51	0.19
	PHCU	1301	0.75	0.10	1284	0.79	1287	0.73	1284	0.47	0.17
	Mature	1250	0.75	0.10	1235	0.78	1238	0.72	1235	0.46	0.16
	New	51	0.89	0.09	49	0.91	49	0.89	49	0.64	0.16
	FHU-A	619	0.82	0.08	609	0.85	610	0.79	609	0.51	0.18
	Mature	531	0.81	0.07	521	0.83	521	0.77	521	0.49	0.16
	New	88	0.93	0.06	88	0.95	89	0.91	88	0.67	0.19
	FHU-B	533	0.89	0.06	524	0.91	524	0.84	524	0.62	0.20
	Mature	476	0.88	0.06	468	0.91	468	0.83	468	0.60	0.20
	New	57	0.95	0.04	56	0.96	56	0.95	56	0.77	0.16
Subsample based on LHUs	Overall scores	2453	0.81	0.10	2417	0.83	2421	0.78	2417	0.48	0.21
	PHCU	1301	0.77	0.10	1284	0.80	1287	0.75	1284	0.46	0.21
	Outside LHUs	1033	0.77	0.09	1021	0.79	1024	0.73	1021	0.40	0.18
	Within LHUs	268	0.80	0.12	263	0.85	263	0.80	263	0.68	0.16
	FHU-A	619	0.83	0.07	609	0.85	610	0.78	609	0.46	0.18
	Outside LHUs	563	0.82	0.07	553	0.84	554	0.77	553	0.43	0.16
	Within LHUs	56	0.88	0.07	56	0.90	56	0.87	56	0.73	0.15
	FHU-B	533	0.89	0.06	524	0.90	524	0.84	524	0.57	0.20
	Outside LHUs	504	0.89	0.05	495	0.90	495	0.83	495	0.56	0.19
	Within LHUs	29	0.91	0.07	29	0.94	29	0.92	29	0.81	0.15
Subsample based on size	Overall scores	2453	0.86	0.10	2417	0.88	2421	0.84	2417	0.53	0.21
	PHCU	1301	0.82	0.12	1284	0.85	1287	0.80	1284	0.50	0.20
	Small	496	0.75	0.12	482	0.80	484	0.73	482	0.51	0.20
	Medium	531	0.85	0.09	529	0.86	529	0.83	529	0.42	0.16
	Large	274	0.89	0.08	273	0.90	274	0.88	273	0.61	0.20
	FHU-A	619	0.88	0.07	609	0.89	610	0.85	609	0.52	0.20
	Small	103	0.86	0.09	103	0.89	104	0.82	103	0.66	0.20
	Medium	462	0.88	0.06	453	0.89	453	0.85	453	0.47	0.18
	Large	54	0.94	0.04	53	0.95	53	0.93	53	0.62	0.19
	FHU-B	533	0.92	0.05	524	0.93	524	0.89	524	0.65	0.19
	Small	11	0.90	0.09	11	0.93	11	0.86	11	0.80	0.17
	Medium	417	0.91	0.05	412	0.92	412	0.88	412	0.63	0.18
	Large	105	0.96	0.04	101	0.97	101	0.95	101	0.69	0.22

References

Abowd, J. M., Kramarz, F., Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica,* 67(2), 251-333.

ACSS. (2014). *Bilhete de Identidade dos Indicadores de Contratualização*. Retrieved August 27, 2019, from http://www2.acss.min-saude.pt/Portals/0/bilhete_identidade_indicadores_contratualizacao_2014____2014_02_25.pdf.

Amado, C. A., Dyson, R. G. (2008). On comparing the performance of primary care providers. *European Journal of Operational Research,* 185(3), 915-932.

Amado, C. A., Dyson, R. G. (2009). Exploring the use of DEA for formative evaluation in primary diabetes care: An application to compare English practices. *Journal of the Operational Research Society,* 60, 1469-1482.

Amado, C. A., Santos, S. (2009). Challenges for performance assessment and improvement in primary health care: The case of the Portuguese health centres. *Health Policy,* 91(1), 43-56.

Andersen, P., Petersen, N. C. (1993). A Procedure for Ranking Efficient Units in Data Envelopment Analysis. *Management Science,* 39(10), 1261-1264.

Aragon, Y., Daouia, A., Thomas-Agnan, C. (2005). Nonparametric Frontier Estimation: A Conditional Quantile-Based Approach. *Econometric Theory,* 21(2), 358-389.

Atkinson, S., Haran, D. (2005). Individual and district scale determinants of users' satisfaction with primary health care in developing countries. *Social Science and Medicine,* 60(3), 501-513.

Atun, R. (2004). *What are the advantages and disadvantages of restructuring a health care system to be more focused on primary care services?* from WHO Regional Office for Europe's Health Evidence Network (HEN).

Banker, R., Charnes, A., Cooper, W. W. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science,* 30(9), 1078-1092.

Barros, P. P. (2017). Competition policy for health care provision in Portugal. *Health Policy,* 121(2), 141-148.

Barros, P. P., Machado, S. M., Simões, J. A. (2011). *Portugal: Health System Review* from The European Observatory on Health Systems and Policies, 13(4).

Barros, P. P., Martinez-Giralt, X. (2003). Preventive Health Care and Payment Systems. *Topics in Economic Analysis & Policy,* 3(1), Article 10.

Battese, G. E., Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics,* 20(2), 325-332.

Cazals, C., Florens, J.-P., Simar, L. (2002). Nonparametric frontier estimation: a robust approach. *Journal of Econometrics,* 106(1), 1-25.

Charnes, A., Cooper, W. W., Golany, B., Seiford, L., Stutz, J. (1985). Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of Econometrics,* 30(1), 91-107.

Charnes, A., Cooper, W. W., Lewin, A. Y., Seiford, L. M. (2013). *Data envelopment analysis: Theory, methodology, and applications*: Springer Science & Business Media.

Charnes, A., Cooper, W. W., Rhodes, E. (1978). Measuring efficiency of decision making units. *European Journal of Operational Research,* 2, 429-444.

Chilingerian, J. A. (1995). Evaluating physician efficiency in hospitals: a multivariate analysis of best practices. *European Journal of Operational Research,* 80(3), 548-574.

Colombo, M. G., Delmastro, M. (2002). The determinants of organizational change and structural inertia: Technological and organizational factors. *Journal of Economics & Management Strategy,* 11(4), 595-635.

Conover, W. (1999). *Practical nonparametric statistics*: John Wiley & Sons. INC, New York.

Cooper, W. W., Seiford, L., Tone, K. (2006). *A Comprehensive Text with Models, Applications, References and DEA-Solver Software* (Second ed.): Springer Science & Business Media.

Cooper, W. W., Seiford, L. M., Zhu, J. (2011). *Handbook on data envelopment analysis* (Second ed. Vol. 164): Springer Science & Business Media.

Cordero-Ferrera, J. M., Cebada, E., Murillo Zamorano, L. (2014). The effect of quality and socio-demographic variables on efficiency measures in primary health care. *European Journal of Health Economics,* 15(3), 289-302.

Cordero-Ferrera, J. M., Crespo-Cebada, E., Murillo-Zamorano, L. R. (2011). Measuring Technical Efficiency in Primary Health Care: The Effect of Exogenous Variables on Results. *Journal of Medical Systems,* 35(4), 545-554.

Cordero, J. M., Alonso-Morán, E., Nuño-Solinis, R., Orueta, J. F., Arce, R. S. (2015). Efficiency assessment of primary care providers: A conditional nonparametric approach. *European Journal of Operational Research,* 240(1), 235-244.

Daraio, C., Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications* (Vol. 4): Springer Science & Business Media.

De Witte, K., Marques, R. C. (2010). Influential observations in frontier models, a robust non-oriented approach to the water sector. *Annals of Operations Research,* 181(1), 377-392.

Deidda, M., Lupiáñez-Villanueva, F., Codagnone, C., Maghiros, I. (2014). Using Data Envelopment Analysis to Analyse the Efficiency of Primary Care Units. *Journal of Medical Systems,* 38(10).

Despacho-4586-A, Ministérios-Finanças/Saúde; (2013).

Despacho-6080-B, Ministérios-Finanças/Saúde; (2014).

Despacho-9999, Ministérios-Finanças/Saúde; (2012).

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal,* 15(1), 292-300.

Dusheiko, M., Goddard, M., Gravelle, H., Jacobs, R. (2008). Explaining trends in concentration of healthcare commissioning in the english NHS. *Health Economics,* 17(8), 907-926.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research,* 132(2), 245-259.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of Royal Statistical Society* 120(3), 253-290.

Ferreira, C., Marques, R. C., Nicola, P. (2013). On evaluating health centers groups in Lisbon and Tagus Valley: efficiency, equity and quality. *Bmc Health Services Research,* 13(529).

Fialho, A. S., Oliveira, M. D., Sa, A. B. (2011). Using discrete event simulation to compare the performance of family health unit and primary health care centre organizational models in Portugal. *Bmc Health Services Research,* 11(274).

Fried, H., Schmidt, S., Yaisawarng, S. (1999). Incorporating the operating environment into a nonparametric measure of technical efficiency. *Journal of Productivity Analysis,* 12, 249-267.

Fried, H. O., Lovell, C. A. K., Schmidt, S. S. (2008). *The measurement of productive efficiency and productivity growth*: Oxford University Press.

Friedberg, M. W., Hussey, P. S., Schneider, E. C. (2010). Primary Care: A Critical Review Of The Evidence On Quality And Costs Of Health Care. *Health Affairs,* 29(5), 766-772.

Giuffrida, A. (1999). Productivity and efficiency changes in primary care: a Malmquist index approach. *Health Care Management Science,* 2(1), 11-26.

Giuffrida, A., Gravelle, H. (2001). Measuring performance in primary care: Econometric analysis and DEA. *Applied Economics,* 33(2), 163-175.

Gouveia, M., Silva, S. N., Oliveira, P., Miguel, L. S. (2006). *Análise dos Custos dos Centros de Saúde e do Regime Remuneratório Experimental* from Grupo de Trabalho da Associação Portuguesa de Economia da Saúde.

Grossman, M. (1972). On the Concept of Health Capital and Demand for Health. *Journal of Political Economy,* 80(2), 223-255.

Hannan, M. T., Freeman, J. (1984). Structural Inertia and Organizational-Change. *American Sociological Review,* 49(2), 149-164.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics,* 17, 1107-1128.

Hollingsworth, B., Dawson, P. J., Maniadakis, N. (1999). Efficiency measurement of health care: a review of non-parametric methods and applications. *Health Care Management,* 2, 161-172.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika,* 30(2), 179–185.

Jürges, H., Pohl, V. (2012). Medical guidelines, physician density, and quality of care: Evidence from German SHARE data. *European Journal of Health Economics,* 13(5), 635-649.

Kringos, D. S., Boerma, W., van der Zee, J., Groenewegen, P. (2013). Europe's strong primary care systems are linked to better population health but also to higher health spending. *Health Affairs,* 32(4), 686-694.

Kringos, D. S., Boerma, W. G. W., Bourgueil, Y., Cartier, T., Hasvold, T., Hutchinson, A., . . . Wilm, S. (2010a). The european primary care monitor: structure, process and outcome indicators. *Bmc Family Practice,* 11, 81.

Kringos, D. S., Boerma, W. G. W., Hutchinson, A., van der Zee, J., Groenewegen, P. P. (2010b). The breadth of primary care: a systematic literature review of its core dimensions. *Bmc Health Services Research,* 10, 65.

Lee, K.-R., Leem, B.-I., Lee, C. W., Lee, C. (2011). Malmquist Productivity Index using DEA frontier in Stata. *The Stata Journal.*

Macinko, J., Dourado, I., Aquino, R., Bonolo Pde, F., Lima-Costa, M. F., Medina, M. G., . . . Turci, M. A. (2010). Major expansion of primary care in Brazil linked to decline in unnecessary hospitalization. *Health Affairs,* 29(12), 2149-2160.

Macinko, J., Starfield, B., Shi, L. (2003). The Contribution of Primary Care Systems to Health Outcomes within Organization for Economic Cooperation and Development (OECD) Countries, 1970–1998. *Health Services Research* 38(3), 831-865.

Mariñoso, B. G., Jelovac, I. (2003). GPs' payment contracts and their referral practice. *Journal of Health Economics,* 22(4), 617-635.

Martin, S., Rice, N., Smith, P. C. (2008). Does health care spending improve health outcomes? Evidence from English programme budgeting data. *Journal of Health Economics,* 27(4), 826-842.

Miguel, L. S., Sá, A. B. (2010). Cuidados de Saúde Primários em 2011-2016: reforçar, expandir Alto Comissariado da Saúde, Ministério da Saúde.

Ministério-Saúde. (2013). *Relatório Anual sobre o Acesso a Cuidados de Saúde no SNS*. Retrieved August 11, 2019, from https://www.sns.gov.pt/wp-content/uploads/2016/07/12072013_RA_Acesso_2012.pdf.

Mooney, G. (1989). The demand for effectiveness, efficiency and equity of health care. *Theoretical Medicine,* 10(3), 195-205.

Murillo-Zamorano, L. R., Petraglia, C. (2011). Technical efficiency in primary health care: does quality matter? *European Journal of Health Economics,* 12(2), 115-125.

OECD. (2013). *Health at a Glance 2013: OECD Indicators*. Retrieved August 28, 2019, from http://dx.doi.org/10.1787/health_glance-2013-en.

OECD. (2015). *Health at a Glance 2015: OECD Indicators*. Retrieved August 28, 2019, from http://dx.doi.org/10.1787/health_glance-2015-en.

OECD. (2018). *Health at a Glance Europe 2018: State of Health in the EU Cycle*. Retrieved August 28, 2019, from https://www.oecd-ilibrary.org/content/publication/health_glance_eur-2018-en.

Olsen, K. R., Gyrd-Hansen, D., Sørensen, T. H., Kristensen, T., Vedsted, P., Street, A. (2013). Organisational determinants of production and efficiency in general practice: A population-based study. *European Journal of Health Economics,* 14(2), 267-276.

Ozcan, Y. A. (2014). *Health Care Benchmarking and Performance Evaluation: An Assessment using Data Envelopment Analysis (DEA)* (Second ed.). Springer Science & Business Media, New York.

Pääkkönen, J., Seppälä, T. T. (2014). Using composite indicators to evaluate the efficiency of health care system. *Applied Economics,* 46(19), 2242-2250.

Pelone, F., Kringos, D. S., Romaniello, A., Archibugi, M., Salsiri, C., Ricciardi, W. (2014). Primary Care Efficiency Measurement Using Data Envelopment Analysis: A Systematic Review. *Journal of Medical Systems,* 39(1).

Pisco, L. (2011). Primary Healthcare Reform in Portugal on two fronts: autonomous family healthcare units and management of groupings of Health Centers. *Ciências saúde coletiva - Rio de Janeiro June 2011; ,* 16(6), 2841-2852.

Preyra, C., Pink, G. (2006). Scale and scope efficiencies through hospital consolidations. *Journal of Health Economics,* 25(6), 1049-1068.

Puig-Junoy, J., Ortún, V. (2004). Cost efficiency in primary care contracting: A stochastic frontier cost function approach. *Health Economics,* 13(12), 1149-1165.

Rocha, F. F., Nishijima, M., Peixoto, S. G. D. (2013). Primary health care policies: Investigation on morbidity. *Applied Economics Letters,* 20(11), 1046-1051.

Rosen, S. (1981). The Economics of Superstars. *American Economic Review,* 71(5), 845-858.

Simar, L. (2003). Detecting outliers in frontier models: A simple approach. *Journal of Productivity Analysis,* 20(3), 391-424.

Simar, L., Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics,* 136(1), 31-64.

Sousa, M. D. C. S., Stošić, B. (2005). Technical Efficiency of the Brazilian Municipalities: Correcting Nonparametric Frontier Measurements for Outliers. *Journal of Productivity Analysis,* 24(2), 157-181.

Staat, M. (2011). Estimating the efficiency of general practitioners controlling for case mix and outlier effects. *Empirical Economics,* 40(2), 321-342.

Starfield, B. (2009). Primary care and equity in health: the importance to effectiveness and equity of responsiveness to people's needs. *Humanity & Society,* 33(February/May), 56-73.

Starfield, B., Shi, L., Macinko, J. (2005). Contribution of Primary Care to Health Systems and Health. *The Milbank Quarterly,* 83(3), 457–502.

Tauchmann, H. (2012). Partial frontier efficiency analysis. *Stata Journal,* 12(3), 461-478.

Torgersen, A. M., Førsund, F. R., Kittelsen, S. A. C. (1996). Slack-adjusted efficiency measures and ranking of efficient units. *Journal of Productivity Analysis,* 7(4), 379-398.

Tribunal-Contas. (2014). *Auditoria ao Desempenho de Unidades Funcionais de Cuidados de Saúde Primários* Retrieved March 6, 2017, from https://erario.tcontas.pt/pt/actos/rel_auditoria/2014/2s/audit-dgtc-rel017-2014-2s.pdf.

Woodcock, S. D. (2008). Wage differentials in the presence of unobserved worker, firm, and match heterogeneity. *Labour Economics,* 15(4), 772-794.

Zhu, J., Cook, W. D. (2007). *Modeling data irregularities and structural complexities in data envelopment analysis*: Springer Science & Business Media.