Designing Incentives for Multitasking Agents: Evidence from Payments to Physicians in England*

Filippo Paternollo† Pietro Tebaldi‡ Andre Veiga§

October 2024

Preliminary, Do Not Circulate or Cite

Abstract

We develop a tractable econometric model in which agents receive rewards when achieving measurable goals across a variety of tasks. Different tasks can interact with others, and the principal (or social planner) would design suboptimal contracts if multitasking was not accounted for. Preferences and costs are identified from variation in contracts and in exposure to different tasks. We apply this framework to study England's "Quality and Outcomes Framework" during the 2009-2019 period, under which thousands of groups of primary care physicians received payments rewarding different clinical outcomes among their patients. We estimate the distribution of physicians' preferences and the degree to which different clinical outcomes are not independent, and then use our model to quantify the impact of the program and to investigate the design of optimal incentives. Our results imply that the program led to a significant improvement in health as measured by increases in quality-adjusted-life-years, net of additional medical costs. We find that re-optimizing the design of the program's incentives would increase net-of-payments health benefits by an additional 3% from baseline.

*We thank helpful comments and suggestions from Matt Gentzkow, Dan Waldinger, Paulo Somaini, Pierre-André Chiappori, Michael Dinnerstein, John Finlay, Simon Lee, Bentley McLeod, and seminar participants at Imperial College London and Stanford University.

†Columbia University. Email: f.paternollo@columbia.edu.

‡Columbia University, and NBER. Email: p.tebaldi@columbia.edu.

§Imperial College London Business School. Email: a.veiga@imperial.ac.uk.

1 Introduction

The design of incentives in principal-agent problems has been central to economic analyses ever since the groundbreaking theoretical contributions in Wilson (1969); Ross (1973); Mirrlees (1976), and Holmström (1979). One decade later, Holmstrom and Milgrom (1991) extended these analyses to instances in which agents must execute multiple tasks that are not independent (effort in one task can raise or lower the cost of effort in another task).[1] More recently, with the development of the *valued-based healthcare* paradigm, agency models gained importance in the design of incentives for medical providers as well (McGuire, 2000; Chalkley and Malcomson, 2000; Einav, Finkelstein and Mahoney, 2018; Gupta, 2021; Gaynor, Mehta and Richards-Shubik, 2023). Still, the gap between theoretical models of multitasking and empirical applications remains substantial and limited to *testing* for multitasking or retrospective policy evaluation.

Prospective and counterfactual analyses of alternative designs of incentives to multitasking agents are often limited by one inherent empirical challenge. In models that include unobserved heterogeneity between agents and multitasking, measuring the cost interactions between different tasks from (exogenous) variation in contracts alone becomes increasingly demanding. With two tasks (e.g. diabetes control and hypertension control) and two contracts, each rewarding one task, the cost model contains not two but *three* parameters: the cost of each task alone, plus one parameter for the interaction between the two (e.g. how likely it is to control diabetes when controlling hypertension for all patients who have both). In large applications in which—as it is often the case—one can only observe limited policy variation that applies at the same time to all agents, the model is likley to be under-identified.

In this article, we show that a class of tractable empirical models of multitasking can be identified by combining time variation in incentives common for all agents with longitudinal

[1] This extension was partly inspired by the prevalence of multitasking in many policy-relevant settings, and particularly by the debate surrounding pay-for-performance for teachers in the early 1990s (Hannaway, 1992). The interaction between pay-for-performance and multitasking in education remains an important object of empirical investigation as of today (see e.g. Dinerstein and Opper, 2022, and references therein).

and cross-sectional variation in exposure to incentives across agents. The key idea is that the importance of different tasks varies across agents and over time (exogenously conditional on observables), which rotates the marginal returns of effort for any level of incentives. In the healthcare context, a physician with (exogenously) more diabetic patients relative to hypertense patients is more exposed to incentives rewarding outcomes for the former group (e.g. blood sugar levels) than for the latter (e.g. blood pressure).

We apply our framework to one of the world's largest pay-for-performance schemes for primary care: the English Quality of Outcomes Framework (QOF). Under this incentive program, thousands of primary care practices receive approximately one-fourth of their annual revenues through payments that are directly linked to a number of measurable outcomes related to their patient pool. For instance, a practice will obtain a larger revenue per cardiac patient if that practice's share of cardiac patients whose blood pressure has been recently measured is higher. In this context, accounting for multitasking is critical (see also Rodríguez-Lesmes and Vera-Hernández, 2021). For instance, effort in controlling patients' blood pressure may also help keep blood sugar levels under control. At the same time, increasing efforts on blood pressure monitoring could limit the time devoted to do comprehensive influenza immunizations or ensure drug prescription adherence. Therefore, understanding the nature of multitasking is central to optimize the provider's incentives. Moreover, despite the QOF's size and twenty years of history, there is yet no consensus on its efficacy and cost-effectiveness.

We combine several public data sources obtained from the National Health Service (NHS) to construct a comprehensive dataset of practice characteristics and QOF incentives and payments for the period 2009-2019. Our analysis focuses on 36 clinical indicators that measure outcomes or processes directly linked to health (e.g. therapy and drug adherence).

The data suggests that physicians respond strongly to incentives. First, one feature of the QOF program is that provider revenue is a piece-wise linear (i.e. kinked) function of achievement in each task. The distribution of achievement across practices in each task

shows significant bunching at the kinks of the functions used to calculate QOF payments. Second, average success in a task is correlated with changes in incentives over time, even when controlling flexibly for practice-indicator-specific unobservables.

The data also suggests that multi-tasking is important (the cost of doing one task is affected by outcomes in other tasks). Holding fixed incentives for a specific indicator, and conditional on practice fixed effects and time-varying number of patients, the correlation between outcomes of different indicators is often statistically significant (and typically positive, suggesting the prevalence of complementarities).

We provide a formal proof of identification of our model using two sources of variation. First, there is longitudinal variation in incentives for each task, which affect all providers in the same way. For instance, the rewards for higher achievement in a given task are raised or lowered by the NHS over time. Second, there is variation in the degree to which each provider is exposed to each task. We observe each practice's geographic location, which can be used as exogenous shifter of patients' composition. Simply put, being located near areas with higher-than-average incidence of diabetes increases, ceteris paribus, the share of diabetic patients and, therefore, the relevance of QOF incentives for diabetes-related clinical outcomes. We show that variation in exposure to each task can be combined with variation in incentives to distinguish all parameters of the model without restrictions on whether each pair of tasks are complements or substitutes.

Our linear-quadratic specification for provider utility implies that the gradient of utility is linear in the the provider's unobserved preferences. This allows us to estimate the distribution of physicians' preferences and costs by maximum likelihood. We find that most pairs of tasks are complements, although we also find evidence of some pairs being substitutes. Multitasking patterns are consistent with intuition: improving outcomes along one clinical dimension for a specific diagnostic group lowers the cost of improving other outcomes for the same patients.

Using our estimates, we perform several counterfactual exercises. First, we quantify the

overall impact of incentives for the QOF indicators included in our analysis. We find that removing all incentives would lead to a drop in the average success rate between 10-60% and lower welfare (benefit net-of-cost) by about 5%.

Second, for 18 indicators (out of the 36 we consider), we collected data on the net-of-cost benefit of a successful outcome, as computed internally by the NHS. Among these, we estimate that the ratio between incentive payments and the median (across practice years) of the resulting net-of-cost health benefits is approximately 1:5.

Third, we use our estimates to optimize the incentives to providers, by defining and solving a (numerically tractable) version of the principal's objective. According to our estimates, the optimized pattern of incentives would result in net health benefits 3% larger than under the status quo. Then, focusing on the 18 indicators for which we observe benefit net-of-cost, we find that our optimized incentive scheme would generate additional health improvements worth 3.9 billion GBP, against additional medical cost of 683 million and additional 199 million in incentive payments to primary care physicians.

In terms of existing empirical work on multitasking, our work contributes to the literature by discussing identification and estimation of a structural model with heterogeneous agents and unrestricted interactions between tasks. Kim, Sudhir and Uetake (2022) pursue the same goal in the context of designing incentives for salesforce in microfinance, where they are able to observe a signal of the agents' effort in response to quasi-random incentives. Prior empirical analyses of multitasking have largely focused on testing theories using field or lab experiments (Slade, 1996; Buser and Peter, 2012; Hong, Hossain, List and Tanaka, 2018; Goes, Ilk, Lin and Zhao, 2018; Manthei and Sliwka, 2019). As mentioned above, education studies pioneered the literature evaluating multitasking in large-scale policy-relevant settings. For a rich review of this literature, as well as a recent example discussing many open issues, see Dinerstein and Opper (2022).

Our application speaks directly to a growing literature discussing the benefits, costs, and optimization of incentives to healthcare providers. Choné and Ma (2011) discuss the

interaction between physician non-financial incentives ("altruism") and optimal incentive design. In the context of Medicare or Medicaid in the US, many studies measure physicians' responses to incentives and the causal effects on healthcare supply or patients' health. These include Clemens and Gottlieb (2014); Gupta (2021); Gaynor et al. (2023); Dunn et al. (2024), and Shi (2024). Counterfactual incentive designs have been proposed in Einav et al. (2018); Einav, Finkelstein, Ji and Mahoney (2022); here we focus on the importance of multitasking when physicians are rewarded separately across a number of different tasks.

Lastly, several articles discussed issues that relate more directly to pay-for-performance and multitasking in the context of managed-care and value-based care. Gaynor, Rebitzer and Taylor (2004) and Mullen, Frank and Rosenthal (2010) focus on physicians' incentives inside HMO organizations, Dumont, Fortin, Jacquemet and Shearer (2008) and Li, Hurley, DeCicca and Buckley (2014) measure the effects of pay-for-performance in primary care using a natural experiment in, respectively, Quebec and Ontario (CA), Rodríguez-Lesmes and Vera-Hernández (2021) proposes a test of multitasking in the context of the QOF program, also highlighting complementarities between indicators. Overall, empirical evidence on the pros and cons of pay-for-performance schemes in medical care, and particularly in primary care, remains mixed. For a review of the vast literature discussing specific issues in the context of the QOF program in England we refer the readers to Gillam, Siriwardena and Steel (2012) and related articles.

2 Modelling and Econometric Framework

2.1 Model

A large number of agents, each indexed by i, are assigned several tasks. In our application, agents are physicians' practices, and tasks are clinical goals for their patients. Alternatively, one could think of a firm assigning tasks to employees. Agents are responsible for delivering these tasks within a given time period, and a principal (payer, or social planner) can set up rewards for agents as a function of performance on each task. Formally, j denotes a

specific, deliverable task (e.g. a patient has blood pressure below a certain threshold, where we will primarily focus on our application henceforth), and $n_{it,j}$ is the number of such tasks assigned to agent i in period t. Every agent in period t is assigned to the collection of tasks $n_{it} = \{n_{it,1}, n_{it,2}, \ldots, n_{it,J}\}$.

At the end of period t the payer observes the collection of outcomes $y_{it} = \{y_{it,j}\}_{j \in J}$, measured as the share of successes: for every j

$$y_{it,j} = \frac{\#\text{ successful tasks } j \text{ by agent } i \text{ in } t}{n_{it,j}} \in [0,1]. \tag{1}$$

The principal sets up an incentive system that pays agents as a function of $y_{it,j}$, proportionally to the number of assigned tasks. Formally, agent i receives a transfer in the amount of

$$r_{it} = \sum_j n_{it,j} \times \rho_{t,j}(y_{it,j}). \tag{2}$$

The collection of functions $\rho_{t,j}(\cdot)$ determines incentives across tasks j in every period t. Importantly, incentives also vary across agents since $\partial r_{it}/\partial y_{it,j}$ varies with $n_{it,j}$. An increase in $\rho_{t,j}(\cdot)$ benefits more agents with a large $n_{it,j}$, and vice-versa. In this formulation, potential successes for task j are treated equally, with no distinction of patients or clients within $n_{it,j}$.

Agent i's payoff is the sum of the total revenues r_{it} and a (money-metric) utility function that varies with n_{it} and y_{it}, which depends on a (private) type θ_{it}: $u(y_{it}, n_{it}; \theta_{it})$. The total utility for an agent with type θ_{it} given the system of incentives is then $U_{it} = u(y_{it}, n_{it}; \theta_{it}) + r_{it}$; this is maximized by i when choosing y_{it}.

This formulation assumes that agents' actions can perfectly control the share of successes, a simplification that represents a substantial departure from the theoretical literature on principal-agent problems with risky outcomes and hidden effort (Holmstrom and Milgrom, 1991). Our choice is dictated by the pursuit of a tractable empirical model that (i) encompasses multitasking, (ii) is identified, and (iii) can be feasibly estimated in large-scale applications. For instances in which the number of potential successes $n_{it,j}$ is sufficiently

large, assuming that agents' production decisions can target precisely the share $y_{it,j}$ might not be too concerning. For small values of $n_{it,j}$, or situations in which outcomes remain highly uncertain even conditional on agents' best effort, the model would need to be enriched further by future work.

Importantly, while payments are additively separable across tasks, the agents' payoffs need not be. That is, to model multitasking we do not assume that $\partial^2 u/\partial y_{it,j}\partial y_{it,k} = 0$ when $j \neq k$. Instead, we make the following assumption

Assumption A1: **Linear-quadratic utility**

There is a symmetric, positive-semi-definite matrix Λ*, with generic entry* $\lambda[j,k]$*, such that*

$$u(y_{it}, n_{it}; \theta_{it}) = \sum_j n_{it,j} \left(y_{it,j} \left(\theta_{i,j} - \left(\sum_k y_{it,k} \times \lambda[j,k] \right) \right) \right). \tag{LQ}$$

Assumption A1 puts a linear-quadratic structure on the utility model (this relates to models of complementary choices presented in Berry et al., 2014; Lewbel and Nesheim, 2019). The matrix Λ "regulates" the interactions between tasks. A positive value of $\lambda[j,k]$ implies that for all agents, increasing $y_{it,j}$ lowers the cost of increasing $y_{it,k}$, and vice-versa. If Λ is diagonal utility becomes additively separable across tasks.

Knowledge of Λ is critical to incentives: if tasks are complements, inducing agents to increase outcomes in one task also increases outcomes and respective payments in other tasks. The opposite holds if tasks are substitutes. We next discuss sufficient conditions for identification of the distribution of types θ_{it} and the matrix Λ.

In Appendix E we show that, for any positive definite matrix Λ, there exists a production function such that total cost can be written as a quadratic form of the outputs vector y.

2.2 Identification

Consider a situation in which every agent i is characterized by the set of observables (x_{it}, z_{it}), and unobservables (θ_{it}, ξ_{it}). The variables in x_{it} (e.g. experience or size of i) can affect

simultaneously task assignment (i.e. the distribution of n_{it}) and the utility type θ_{it}. Instead, z_{it} collects instrumental variables (e.g. geographic location) that affect task assignment but not θ_{it}. Lastly, ξ_{it} is unobservable and can potentially affect both task assignment and θ_{it}. When discussing identification of the model, we maintain that the econometrician observes the conditional density $g\left(y_{it}|x_{it}, n_{it}, z_{it}\right)$ over the entire support of (y_{it}, x_{it}, n_{it}), for all t.

We can now complete the list of our identifying assumptions.

Assumption A2: **Shifters of task assignment and conditional independence**
There is a known function σ such that there is a unique value $\widehat{\xi}_{it}$ for which $n_{it} = \sigma(z_{it}, x_{it}, \widehat{\xi}_{it})$, and $\widehat{\xi}_{it} = \xi_{it}$. Moreover, there is a family of conditional densities $\{f_j\left(\cdot|x_{it}, \xi_{it}\right)\}_j$ such that $\theta_{it} \underset{iid}{\sim} \prod_j f_j\left(\theta_{it,j}|x_{it}, \xi_{it}\right)$, and $\theta_{it}|x_{it}, \xi_{it} \perp n_{it}|x_{it}, \xi_{it}$.

Assumption A3: **Variation in incentives and task assignments (completeness)**
Incentive payments are such that $\rho_{t,j}(\cdot) = \rho(\cdot; \alpha_{t,j})$, where the family of functions ρ is differentiable with respect to all arguments. Moreover, for all functions $B(y_{it})$ with finite expectations, if $\mathbb{E}\left[B(y_{it})|n_{it}, x_{it}, \xi_{it}, \alpha_t\right] = 0$ almost surely, then $B(y_{it}) = 0$ almost surely.

Assumption A2 states that variation in the instrumental variables z_{it} can be used to identify the unobservables ξ_{it} affecting n_{it} conditional on x_{it}. This argument is standard to the discrete-choice demand literature and formalized in Berry and Haile (2014). Assumption A2 also requires the distribution of private types to be independent across i and j after conditioning on (x_{it}, ξ_{it}).

Assumption A3 is a completeness (full-rank) assumption. It requires the variation in incentives and exposure to incentives—as determined by changes in n_{it} generated by shifts in z_{it} conditional on x_{it} and ξ_{it}—to be sufficient to distinguish any two functions of the outcomes. This assumption is generally violated in finite samples, and estimation will therefore rely on distributional assumptions on $f_j(\cdot)$. Nevertheless, Assumption A3 serves the important purpose of highlighting conditions in the data-generating process that are sufficient to identify the model's primitives.

We then obtain the following

Proposition 1. *Under A1-A3,* $\{f_j(\cdot|x_{it},\xi_{it})\}_j$ *and* Λ *are identified.*

The proof is in Appendix A, here we discuss the case with two tasks.

First, variation in z_{it} ensures that the unobservables ξ_{it} are identified following the argument in Berry and Haile (2014).[2] In the example of our empirical application, observing a practice that captures an equal or larger share of patients albeit being farther away, ceteris paribus, implies that it must be of unobservably higher quality.

Consider now two tasks, say $j = A, B$. The goal is to identify $f_j(\theta_{it,j}|x_{it},\xi_{it})$ for each j, and the three distinct entries of Λ: $\lambda[A,A], \lambda[B,B]$, and the off-diagonal element $\lambda[A,B] = \lambda[B,A]$ which is the essential parameter that distinguishes our model to one without multitasking. To uniquely determine these parameters from the data, Assumption 3 allows us to consider arbitrary variation in incentives (parametrized by α) and patients' composition $(n_{it,A}, n_{it,B})$ conditional on (x_{it},ξ_{it}). In everything that follows we simplify our discussion and notation omitting the dependence and conditioning on (x_{it},ξ_{it}), without any loss of generality. (See Appendix A for details.)

Consider it pairs for which $n_{it,B} = 0$, so that—differentiating (LQ), and dividing by $n_{it,A}$—$y_{it,A}$ must solve $\theta_{it,A} + \partial\rho(y_{it,A};\alpha_{t,A})/\partial y_{it,A} - 2\lambda[A,A]y_{it,A} = 0$, or

$$\theta_{it,A} = 2\lambda[A,A]y_{it,A} - \frac{\partial\rho(y_{it,A};\alpha_{t,A})}{\partial y_{it,A}}. \tag{3}$$

The log-density of $y_{it,A}$ for these it pairs is then $\log(g(y_{it,A})) = \log(2\lambda[A,A]) + \log\left(f_j(\theta_{it,A})\right)$, where $\theta_{it,A}$ is determined in (3). Since this has full support, we can differentiate it with respect to $y_{it,A}$ and with respect to $\alpha_{t,A}$, and obtain

$$\lambda[A,A] = \frac{1}{2}\left(-\frac{\partial\log(g(y_{it,A}))/\partial y_{it,A}}{\partial\log(g(y_{it,A}))/\partial\alpha_{t,A}}\frac{\partial^2\rho(y_{it,A})}{\partial y_{it,A}^2} + \frac{\partial^2\rho(y_{it,A})}{\partial y_{it,A}\partial\alpha_{t,A}}\right). \tag{4}$$

[2]Under Assumption 2, one can invert σ by finding the value of ξ_{it} that leads to n_{it} holding fixed x_{it} and z_{it}: $\xi_{it} = \sigma^{-1}(n_{it};x_{it},z_{it})$. Conditional on (x_{it},n_{it}), changes in z_{it} must be compensated by changes in ξ_{it}.

While this expression might seem cumbersome, the intuition is as follows. At any point in the data, it must be that marginal revenue equals marginal cost. Under Assumption A1, marginal revenue depends on the unknown type $\theta_{it,A}$ and on the known marginal incentives $\partial\rho(y_{it,A};\alpha_{t,A})/\partial y_{it,A}$, while marginal cost depends instead on the product of $y_{it,A}$ and the unknown parameter $\lambda[A,A]$. Variation in $\alpha_{t,A}$—which appears in the denominator of (4)—identifies (the distribution of private types that leads to) marginal revenue, say $MR_{it}(\alpha_{t,A})$. Indeed, any increase in $\alpha_{t,A}$ must correspond to an increase in $y_{it,A}$ or to a decrease in $\theta_{it,A}$. Then, $\lambda[A,A]$ must be the parameter for which $MC_{it}(\lambda[A,A]) = MR_{it}(\alpha_{t,A})$ for the same distribution of private types. The argument applies symmetrically to identify $\lambda[B,B]$ using it pairs for which $n_{it,A} = 0$.

Then, considering it pairs for which $n_{it,A} \neq 0$ and $n_{it,B} \neq 0$, y_{it} must solve

$$n_{it,A}\left(\theta_{it,A} + \partial\rho(y_{it,A};\alpha_{t,A})/\partial y_{it,A}\right) - 2n_{it,A}\lambda[A,A]y_{it,A} - (n_{it,A}+n_{it,B})\lambda[A,B]y_{it,B} = 0,$$

and similarly for B. Because $\lambda[A,A]$ and $\lambda[B,B]$ can now be treated as known, we can once again use a similar argument. After deriving the (joint) log-density of y_{it}, its local variation relative to variation induced by changes in incentives (α_t)—which identifies the distribution of private types—uniquely determines $\lambda[A,B]$, which is the only parameter affecting marginal cost left to be identified. Once all the parameters in Λ are identified, the distribution of θ_{it} is nonparametrically identified from first-order conditions (e.g. equation (3) when $n_{it,B} = 0$).

Our argument combines the variation in incentives common for all agents with variation in exposure to these incentives. When $n_{it,B} = 0$, only $\alpha_{it,A}$ matters, and vice-versa. The model is generally not identified from variation in α_t alone since its dimension is J while the dimension of Λ is $J(J-1)/2$ and the dimension of θ_{it} is J.

3 Incentives to Physicians in the English National Health Service

3.1 The Quality of Outcomes Framework

The Quality of Outcomes Framework (QOF henceforth) is arguably the most comprehensive national primary care pay-for-performance scheme in the world (see, e.g. Gillam et al., 2012,

and references therein). It applies to virtually all primary care practices—also known as general practitioners—in the English National Health Service (NHS).[3] It was introduced as part of a new national physician fee-for-service contract in 2004. Since then, it endured many changes and remains an object of policy debate. On the one hand, transfers total approximately one billion GBP per year. On the other hand, research has yet to be conclusive that the realized population's health improvements justify the program's costs.

In terms of incentive design (see Appendix B for more details), for every practice-year the QOF measures the share of successes across approximately one hundred indicators,[4] which we index by j as above. At the end of year t, a practice i receives a transfer equal to

$$r_{it} = \sum_j n_{it,j} \underbrace{\max\left\{0, \alpha_{t,j} \times \min\left\{y_{it,j} - \underline{Y}_{t,j}, \overline{Y}_{t,j} - \underline{Y}_{t,j}\right\}\right\}}_{\rho_j(y_{it,j})}. \tag{5}$$

That is, for every indicator j the practice receives nothing if the average achievement $y_{it,j}$ is lower than the "floor" $\underline{Y}_{t,j}$, and then receives $\alpha_{t,j}$ per-relevant-patient for every percentage increase in achievement up until the "ceiling" $\overline{Y}_{t,j}$, after which compensation stops.

One example of these contracts and how these varied over time is illustrated in the top panel of Figure 1. For indicator DM11, which measures the share of diabetic patients for whom a critical marker is below the recommended threshold, we show the function $\rho_j(y_{it,j})$ for $t = 2013, 2014$. This function varies yearly, but we selected these two years for illustrative purposes. While $\underline{Y}_{2013,j} = \underline{Y}_{2014,j}$, the program altered incentives for this indicator by setting $\overline{Y}_{2014,j} > \overline{Y}_{2013,j}$, while $\alpha_{2014,j} < \alpha_{2013,j}$. As a result of these changes, the amount paid in 2013 to a practice with 300 diabetic patients with share of successes 0.6 would have been almost equal to the amount paid in 2014 with a 10 percentage point higher share of successes.

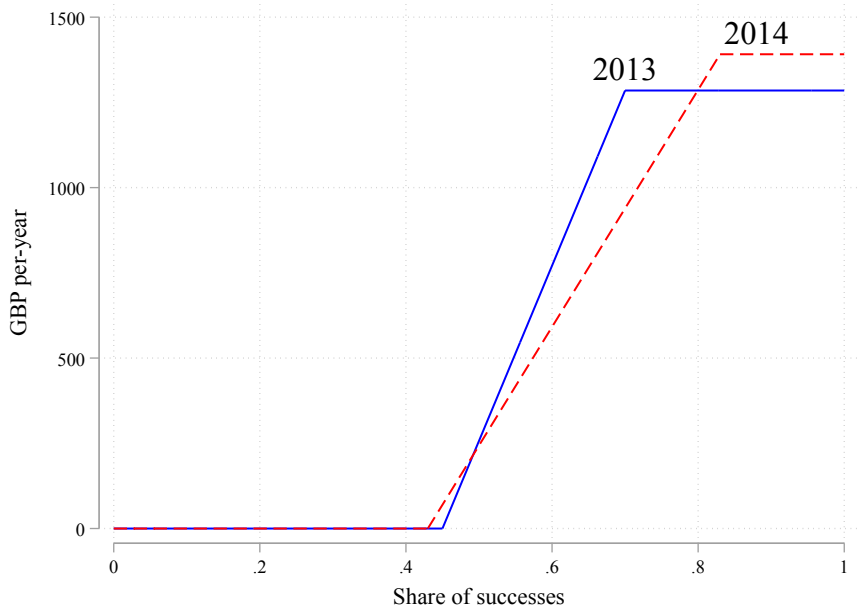
The bottom panel of Figure 1 provides a first, suggestive look at the (seemingly rational)

[3]The program is voluntary, but participation is almost universal. In our analyses, when merging administrative QOF data with other NHS datasets at the practice level, we found only 48 out of 7,935 practices for which QOF data are unavailable.

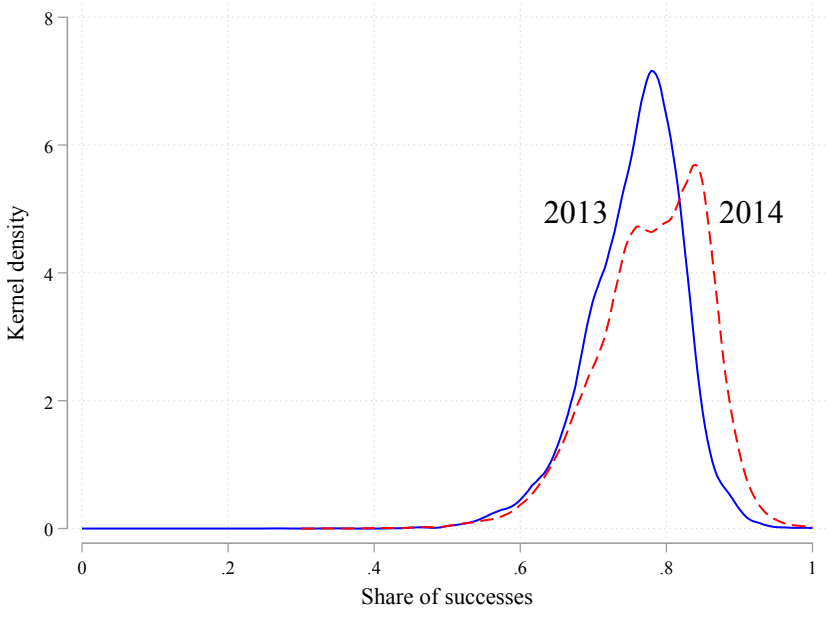
[4]These are grouped in five distinct domains: clinical; public health; additional services; vaccination and immunization; quality improvement. We focus on the clinical domain.

Figure 1: Example of incentives for QOF indicator *DM11*

% diabetes patients in whom the last glycohaemoglobin IFCC-HbA1c is 64 mmol/mol or less.



(a) Payments for a practice with 300 relevant patients



(b) Distribution of share of successes across practices

response of outcomes to changes in incentives, anticipating our discussion below. Plotting the density of share of successes across practices in $t = 2013, 2014$ shows that the distribution shifts to the right as the ceiling and the strength of incentives increase. Moreover, in both years the density peaks right around the kink in the payment function, as already observed in Rodríguez-Lesmes and Vera-Hernández (2021).

We next introduce our data and investigate further the relationship between variation in QOF incentives and the distribution of clinical outcomes across practices.

3.2 Data

We employ data on patient composition, successes, and incentives for QOF indicators belonging to the clinical domain during the 2009-2019 period. We focus on indicators measuring health outcomes, or intermediate outcomes related to health (e.g. drug prescriptions and adherence). This results in a final dataset consisting of 36 indicators during our study period, complemented with information on practice characteristics.

Our analysis dataset is built by combining several public sources obtained directly from the NHS website, as described in Appendix B. This results in 2.2 million observations, each corresponding to a unique combination of practice-year-indicator level, summarized in Table 1.

The top panel of Table 1 highlights the variation in incentives across indicators and years, as well as the benefits and cost of a success (for one patient) as measured by the National Institute for Health and Care Excellence (see also Appendix B). An average value of $\alpha_{t,j}$ equal to 31.07 corresponds to the increase in revenue r_{it} for a practice with 100 patients for which indicator j is relevant, as long as the share of successes is between $\underline{Y}_{t,j}$ and $\overline{Y}_{t,j}$. This incentive parameter varies from less than 10 to more than 90 GBP, and its distribution is skewed to the right with some indicators being rewarded very highly. This skeweness reflects the distribution of the "societal value" of a success, as measured by the increase in QALY for the patient. This is 11099 GBP for the average indicator, with a median value of 4069 GBP,

Table 1: Summary statistics

	Mean	Std. Dev.	P-10	Median	P-90	Obs.
Panel A: Indicator-year						
Indicator-year QOF incentive parameters:						
$\alpha_{t,j}$	31.07	33.73	7.09	18.02	91.01	276
$\overline{Y}_{t,j}$	0.816	0.12	0.65	0.8	0.97	276
$\underline{Y}_{t,j}$	0.446	0.069	0.4	0.4	0.57	276
Value of success for one patient:[(a),(b)]						
Extra QALY	11099	17056	113	4069	45813	142
Cost	2966	7388	6	30	9662	142
Panel B: Indicator-practice-year						
Share of successes $y_{it,j}$	0.874	0.115	0.723	0.9	1	2154236
Number of patients $n_{it,j}$	196	272	10	108	468	2154236
Revenues per patient $\rho_j(y_{it,j})$	9.70	9.15	2.16	6.75	25.19	2154236
Panel C: Practice-year						
QOF outcomes:						
Total revenues r_{it}	33540	24389	9639	27441	65999	85509
N. indicators with $n_{it,j} > 0$	25.2	3.3	22	25	31	85509
Year	2013.8	3.1	2010	2014	2018	85509
Practice characteristics:[(c)]						
Number of physicians	5.26	3.85	1	4	10	81535
Avg. age	46.47	7.98	38.76	45.75	57	81535
Share fully qualified	0.96	0.1	0.8	1	1	81535
Avg. distance to patient (km)	1.83	1.26	0.9	1.51	3.04	37238

Notes: *The table summarizes our analysis dataset, built following the procedure described in Appendix B. Share of successes is defined in equation (1); QOF incentive parameters, revenues per patient—$\rho_j(y_{it,j})$—and total revenues—r_{it}—are defined in equation (5). All monetary values are in 2020 GBP.*

[(a)]: *Value of success is obtained for* 18 *indicators from reports provided by the National Institute for Health and Care Excellence and from the Department of Health. See Appendix B for more details.*

[(b)]: *1 additional QALY is valued 25,000 GBP, the "societal value" used by the National Institute for Health and Care Excellence.*

[(c)]: *Qualified physicians are those who have received all qualifications and are not undergoing training for general practice.*

and a 90th percentile equal to 45813 GBP.[5]

The middle panel of Table 1 summarizes outcomes at the practice-year-indicator level. On average, practices have 196 patients in each indicator, and record a high success rate (0.874) which is 6 percentage points higher than the average level of the ceiling $\overline{Y}_{t,j}$ (0.816).

[5]This corresponds to indicator AF3, described as *In those patients with atrial fibrillation in whom there is a record of a CHADS2 score of 1 (latest in the preceding 15 months), the percentage of patients who are currently treated with anti-coagulation drug therapy or an anti-platelet therapy.*

This leads to revenue per patient in the amount of 9.70, on average, adding up to average annual revenue for 33540 GBP per practice-year (panel C) resulting from the 36 indicators included in our analysis. The modal practice-year has patients in 25 out of 36 indicators.[6]

Panel C of Table 1 also summarizes the characteristics of practices in our sample. These employ, on average, 5.26 physicians, with age distribution centered at 45.75 and the vast majority of whom has the highest qualifications for primary care according to NHS standards. Importantly for our analysis, patients tend to select a practice very close to their residence: the average patient needs to travel only 1.83 kilometers to see their doctor, and for less than ten percent of patients this distance is greater than 3.04.

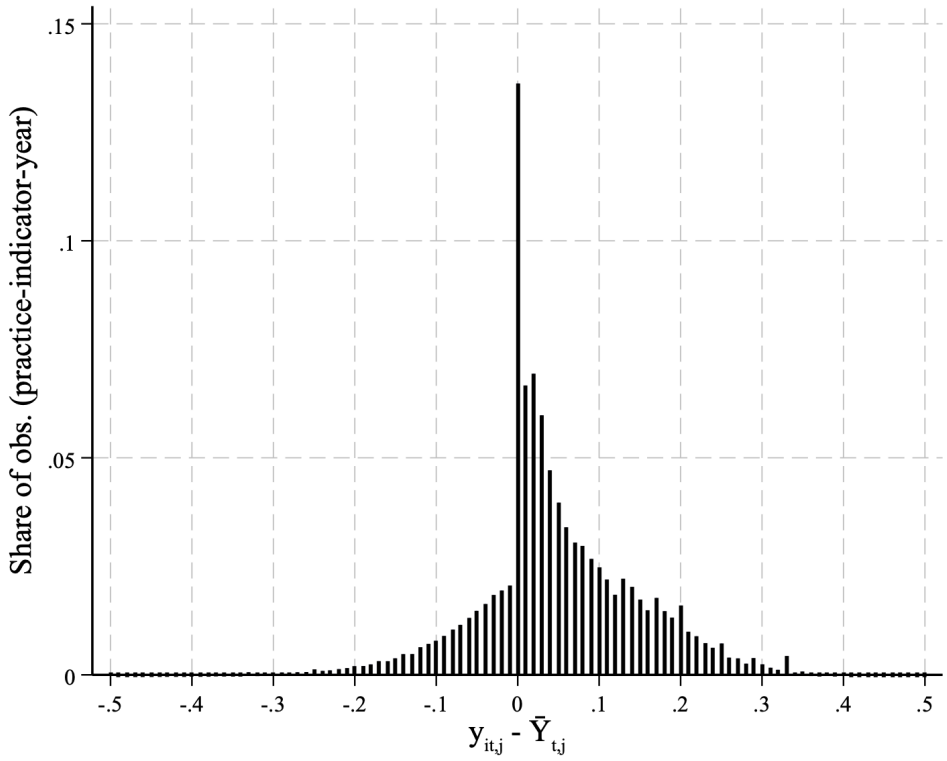
3.3 Empirical Evidence

Bunching. The example presented in Figure 1 above suggested that physician practices seem to respond to QOF incentives. To corroborate this further, we use our data to inspect the distribution of the distance between realized share of successes ($y_{it,j}$) and the ceiling $\overline{Y}_{t,j}$ above which incentives to increase success further for indicator j become null. If physicians were responsive to incentives the distribution of $y_{it,j} - \overline{Y}_{t,j}$ should show significant bunching at 0. Moreover, if physicians were solely motivated by monetary rewards *and* indicators were independent there should be few instances in which $y_{it,j} - \overline{Y}_{t,j} > 0$.

Figure 2 illustrates the distribution of $y_{it,j} - \overline{Y}_{t,j}$ across all itj triplets in our analysis dataset. The prevalence of bunching at the kink of monetary incentives is starking, with frequency greater than 0.13. In addition, it is often the case that the share of successes exceeds the ceiling $\overline{Y}_{t,j}$. This suggests that physicians may be driven by motives other than pure revenue maximization (see also Choné and Ma, 2011; Godager and Wiesen, 2013; Einav et al., 2018; Gaynor et al., 2023, for models of physicians' altruism), or that indicators may be complementary to one another and increasing success in j may lead to higher success in k.

[6]Some indicators are only present in a subset of years, as illustrated in Appendix Figure C.1. Our current analysis does not account for nor leverage this fact, which may be incorporated in future work.

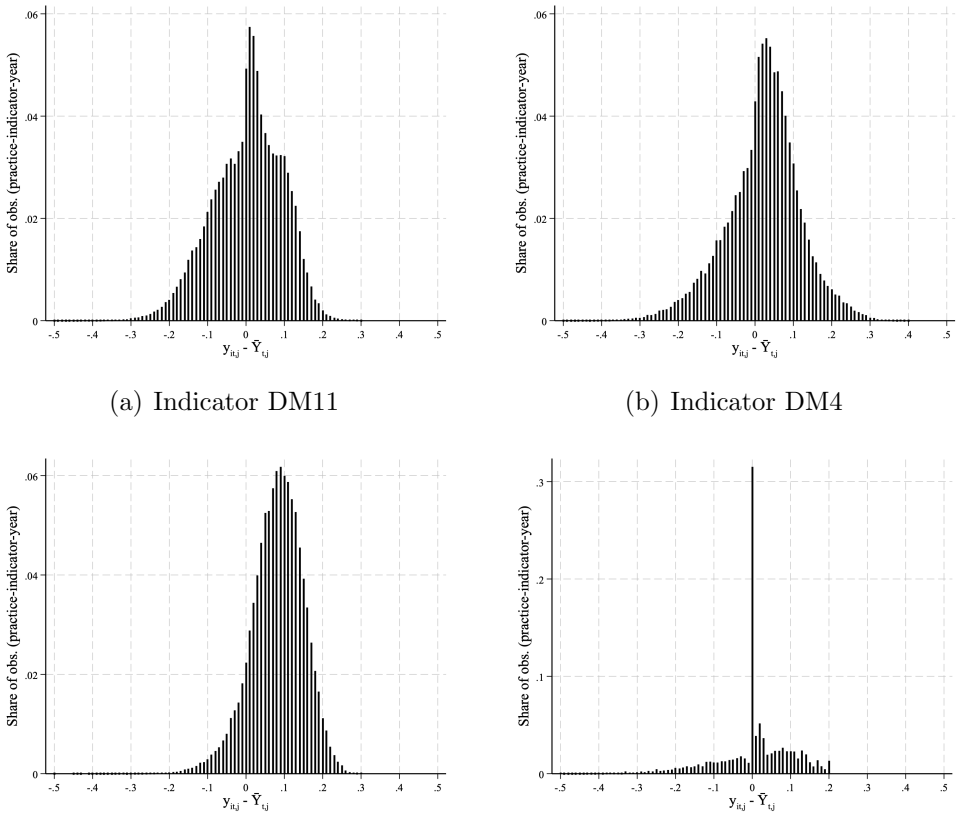
Figure 2: Distribution of "excess" share of successes $y_{it,j} - \overline{Y}_{t,j}$



Notes: *Distribution of the distance between realized share of successes ($y_{it,j}$) and the ceiling $\overline{Y}_{t,j}$; see equation (5). We weight observations by $N_{it} = \sum_j n_{it,j}$. We correct our calculations to account for the fact that the number of patients is integer and it may therefore be impossible for a practice to reach exactly $\overline{Y}_{t,j}$. Therefore, we consider the distance between $y_{it,j}$ and the ratio between the smallest integer larger than $n_{it,j}\overline{Y}_{t,j}$ and $n_{it,j}$. This is the relevant bunching point, since represents the first patient whose success does not lead to additional revenues.*

The aggregate distribution in Figure 2 masks significant heterogeneity across indicators. This is illustrated in Figure 3, where for four indicators relating to diabetic patients. For the two indicators in the top panels, DM11 and DM4, bunching is evident but more limited; the frequency at zero is lower than 0.05 and the "jump" at zero approximately one percentage point. For indicator DM15 in the bottom-left panel there is no evidence of bunching. On the other extreme, for indicator DM8, the frequency at zero is almost 0.4. Perhaps not surprisingly, DM8 is the only indicator out of the four considered in the figure that does not correspond to a clinical outcome, but instead checks the share of patients that are receiving a recommended treatment. One may therefore expect the variance of physicians' heterogeneity in terms of preferences and costs for this indicator to be smaller.

Figure 3: Distribution of "excess" share of successes for specific indicators



(a) Indicator DM11

(b) Indicator DM4

(c) Indicator DM5

(d) Indicator DM8

Notes: *Distribution of the distance between realized share of successes ($y_{it,j}$) and the ceiling $\overline{Y}_{t,j}$; see equation (5) and note to Figure 2 for references. We weigh observations by $N_{it} = \sum_j n_{it,j}$. Indicator DM11:* The percentage of patients with diabetes, on the register, in whom the last IFCC-HbA1c is 64 mmol/mol or less in the preceding 12 months. *Indicator DM4:* The percentage of patients with diabetes, on the register, in whom the last blood pressure reading (measured in the preceding 12 months) is 140/80 mmHg or less. *Indicator DM5:* The percentage of patients with diabetes whose last measured total cholesterol within the preceding 15 months is 5mmol/l or less. *Indicator DM8:* The percentage of patients with diabetes, on the register, with a diagnosis of nephropathy (clinical proteinuria) or micro-albuminuria who are currently treated with an ACE-I (or ARBs).

Multitasking. We now explore the interaction between different indicators, and the relationship between QOF incentives, share of successes, and the unobservables affecting the sorting of patients across practices. To provide empirical support for our emphasis on multitasking and interactions between indicators, we follow an intuition related to the test suggested by Rodríguez-Lesmes and Vera-Hernández (2021). For every indicator j, we control for practice identifier, j-specific QOF incentive parameters $(\alpha_{t,j}, \underline{Y}_{t,j}, \overline{Y}_{t,j})$, and regress the share of successes $y_{it,j}$ exceeding $\underline{Y}_{t,j}$ on the collection of QOF incentive parameters

$(\alpha_{t,k}, \underline{Y}_{t,k}, \overline{Y}_{t,k})$ of other indicators $k \neq j$. Specifically, we estimate

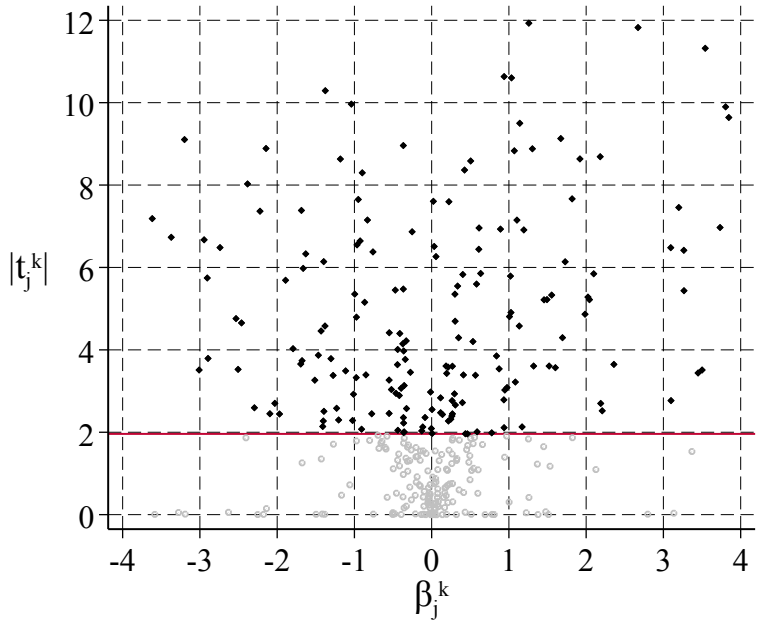
$$y_{it,j} - \underline{Y}_{t,j} = \sum_{k \neq j} \beta_j^k (\alpha_{t,k} \times (\overline{Y}_{t,k} - \underline{Y}_{t,k})) + \gamma_j (\alpha_{t,j} \times (\overline{Y}_{t,j} - \underline{Y}_{t,j})) + \mu_{i,j} + \varepsilon_{it,j}. \quad (6)$$

A limitation of this approach is that it solely relies on the variation in the QOF design variables (see Figure C.1). This makes it impossible to estimate all the relevant coefficients due to collinearity issues. We estimate 426 out of a total of 1260 possible estimable β_j^k coefficients. In Section 4, we overcome these challenges by developing a model that enables us to identify the dependencies between indicators, leveraging the full amount of identifying variation in our data.

Figure 4 illustrates our estimates from equation (6). Coefficients significantly different from zero suggest dependencies between indicators: if positive, indicators are complements, if negative, indicators are substitutes. We plot in black the significant coefficients, plotted in the space of their size (horizontal axis) and t-statistics in absolute value (vertical axis). The shaded coefficients are not significantly different from zero. This shows that there are many pairs of indicators (243 out of 426; 58%) for which this simple test rejects the null hypotheses that they are independent, suggesting that—in the language of our model—$\lambda[j,k] \neq 0$. Inspecting the signs, these estimates suggest that significant complementarities (30%) might be more prominent than substitutabilities (28%).

Distance and practice choice. To estimate our model we need to isolate variation in incentives from unobserved heterogeneity that may affect the physicians' utility. On the one hand, we can exploit directly the variation in QOF design over time (see also Appendix Figure C.1). On the other hand, without a very (long panel and) rich variation in $\alpha_{t,j}$ across t and j, our discussion of identification in Section 2 points at variation in n_{it} across i and t as a key source of variation. Combining changes in n_{it} with changes in QOF parameters leads to rich variation in the intensity of incentives for different indicators across practices

Figure 4: Coefficient estimates and t-statistics



Notes: Estimated coefficients and t-statistics (in absolute value) from J regressions of $y_{it,j} - \underline{Y}_{t,j}$ on $\alpha_{t,k} \times (\overline{Y}_{t,k} - \underline{Y}_{t,k})$ for $k \in J \setminus \{j\}$, controlling for time-varying policy parameters $\overline{Y}_{t,j}, \underline{Y}_{t,j}$ and $\alpha_{t,j}$.

and over time as illustrated for two indicators in Appendix Figure C.2.

However, a possible concern is that patients sort along dimensions that affect outcomes (see also Brown, Hansman, Keener and Veiga, 2023). To isolate exogenous variation in patient composition we use variation in practice location, leading to variation in distance between i and potential patients associated with different indicators. We use standard techniques widely used in empirical industrial organization for the analysis of discrete-choice models, where our goal is to estimate unobservables that affect patients' sorting between practices and may be correlated with private types and hence QOF outcomes.

To account for this, we estimate a model of practice choice by patients, relying on data on patient composition across geographic areas. If a patient for which indicator j is relevant is willing to travel farther to be registered with physician i in year t, this is captured by a larger value of $\xi_{it,j}$. We recover estimates of $\xi_{it,j}$ for all i, j, and t, and then allow these to affect outcomes. Specifically, in what follows, we will allow $\widehat{\xi}_{it,j}$ to affect the distribution of

the practice type (θ_i), and therefore clinical outcomes. In Appendix Figure C.4 we show that $\widehat{\xi}_{it,j}$ is indeed positive correlated with $y_{it,j}$, even after controlling for indicator fixed-effects.[7]

Incentives and share of successes. Lastly, in Table 2 we summarize our preliminary finding estimating alternative regressions investigating the response of success rate to QOF incentives and a practice's share of relevant patients $\left(\frac{n_{it,j}}{\sum_j n_{it,j}}\right)$, also controlling for practice characteristics. Specifically, we estimate

$$y_{it,j} - \underline{Y}_{t,j} = \beta_1 \left(\alpha_{t,j} \times (\overline{Y}_{t,j} - \underline{Y}_{t,j})\right) + \beta_2 \left(\frac{n_{it,j}}{\sum_j n_{it,j}}\right) + \beta_3 \left(\frac{n_{it,j}}{\sum_j n_{it,j}} \times \alpha_{t,j} \times (\overline{Y}_{t,j} - \underline{Y}_{t,j})\right) + \gamma X_{it} + [\kappa_j][+\nu_i] + \varepsilon_{it,j} \tag{7}$$

where for some specifications we include indicator fixed-effects (κ_j) or practice fixed-effects (ν_i). Once again, we leverage each practice's geographic location as an exogenous shifter of patients' composition and instrument a practice's share of relevant patients with the share of relevant patients in practices within a 10km radius.

The impact of QOF incentives as set by $\alpha_{t,j} \times (\overline{Y}_{t,j} - \underline{Y}_{t,j})$ is positive, significant, and robust across specifications. The effect of the share of relevant patients becomes positive once we address endogeneity concerns and instrument it with the composition of patients in nearby practices. Moreover, the coefficient associated with the interaction between QOF incentives and the share of relevant patients is positive and significant, indicating that the effect of QOF incentives on a practice's success rate is increasing in that practice's exposure to that specific indicator.

[7]Since the data needed for demand estimation starts in 2015 but our main data starts in 2009, we inpute the estimated demand residual $\widehat{\xi}_{it,j}$ for the years before 2015. When possible, we take the first available value for a given practice-indicator-year. For indicators present only before 2015, we take the average within the practice and indicator-group for the post-2014 period. If the entire indicator group is available only before 2015, we take the practice-level average for the post-2014 period.

Table 2: Determinants of QOF outcomes

	Dependent variable: $y_{it,j} - \underline{Y}_{t,j}$				
	OLS	OLS	OLS	OLS	IV
	(1)	(2)	(3)	(4)	(5)
$\alpha_{t,j} \times (\overline{Y}_{t,j} - \underline{Y}_{t,j})$	0.117	0.278	0.302	0.289	0.24
	(0.001)	(0.003)	(0.003)	(0.004)	(0.004)
$\frac{n_{it,j}}{\sum_j n_{it,j}}$		−0.541	−0.443	−0.425	0.084
		(0.009)	(0.008)	(0.009)	(0.011)
$\frac{n_{it,j}}{\sum_j n_{it,j}} \times \alpha_{t,j} \times (\overline{Y}_{t,j} - \underline{Y}_{t,j})$		3.008	1.62	1.684	3.152
		(0.091)	(0.089)	(0.091)	(0.109)
Controls			Yes	Yes	Yes
FE		Ind.	Ind., Practice	Ind., Practice	Ind., Practice
R-squared	0.012	0.285	0.362	0.363	-
Observations	2145595	2145595	2145595	2014257	2005257

Notes: We weigh observations by $N_{it} = \sum_j n_{it,j}$. We divide $\alpha_{t,j} \times (\overline{Y}_{t,j} - \underline{Y}_{t,j})$ by 10 both in the main effects and in the interaction terms. The controls include the practice's number of practitioners, average practitioner age, and share of fully qualified practitioners. Robust standard errors in parentheses.

4 Estimating Preferences of Multitasking Physicians

4.1 Parametric Specification and Estimation

We next apply the identification strategy discussed in Section 2 in our empirical context. This allows us to parametrically estimate the distribution of physicians' types and the cost matrix that characterizes the interactions between QOF indicators.

For every practice i we consider the set of observables $\widetilde{x}_i = \left(x_i, \widehat{\xi}_i\right)$, where x_i collects a constant term, practice size, average age of physicians, and share of fully qualified physicians, while $\widehat{\xi}_i$ is the estimated unobservable we previously estimated. We assume that, for every i, t, and j

$$\theta_{it,j} \underset{iid}{\sim} \mathcal{N}\left(\mu_j \widetilde{x}_i, \sigma_j\right), \quad \text{and} \quad n_{it} | \widetilde{x}_i \perp \theta_{it} | \widetilde{x}_i. \tag{8}$$

Then, imposing LQ-utility as specified in Assumption 1, the parameters of the model consists

of the matrix Λ and the collection $\{\mu_j, \sigma_j\}_j$.

These can be estimated via maximum likelihood, after noticing that, for a given realization of y_{it}, first-order optimality implies that

$$\begin{aligned}
&n_{it,j}\theta_{it,j} \leq \nabla_{it,j}(y_{it}) && \text{if } \; y_{it,j} = 0 \\
&n_{it,j}\theta_{it,j} = \nabla_{it,j}(y_{it}) && \text{if } \; y_{it,j} \in (0, \underline{Y}_{t,j}], \text{ or } y_{it,j} \in (\overline{Y}_{t,j}, 1) \\
&n_{it,j}\left(\theta_{it,j} + \alpha_{t,j}\right) = \nabla_{it,j}(y_{it}) && \text{if } \; y_{it,j} \in (\underline{Y}_{t,j}, \overline{Y}_{t,j}) \\
&n_{it,j}\theta_{it,j} \leq \nabla_{it,j}(y_{it}) \leq n_{it,j}\left(\theta_{it,j} + \alpha_{t,j}\right) && \text{if } \; y_{it,j} = \overline{Y}_{t,j} \\
&n_{it,j}\theta_{it,j} \geq \nabla_{it,j}(y_{it}) && \text{if } \; y_{it,j} = 1,
\end{aligned} \tag{9}$$

where we set

$$\nabla_{it,j}(y_{it}) \equiv \sum_{k \in J} \left(n_{it,j} + n_{it,k}\right) y_{it,k} \times \lambda[j,k].$$

The likelihood of y_{it} is then derived analytically as the distribution of a discrete-continuous variable. Formally, letting B_{it}^0 be the set of indicators j that are bunched at $y_{it,j} = 0$ for the pair it, and similarly for $B_{it}^{\overline{Y}}$ and B_{it}^1, and denoting with $\Phi(\cdot; a, b)$ the CDF of a Gaussian distribution with mean a and standard deviation b (with PDF $\phi(\cdot; a, b)$), the likelihood is

$$\begin{aligned}
\ell_{it} = &\prod_{j \in B_{it}^0} \Phi(\nabla_{it,j}(y_{it}); n_{it,j}\mu_j\widetilde{x}_i, n_{it,j}\sigma_j) \\
&\times \prod_{j \in B_{it}^{\overline{Y}}} \left(\Phi(\nabla_{it,j}(y_{it}); n_{it,j}\mu_j\widetilde{x}_i, n_{it,j}\sigma_j) - \Phi(\nabla_{it,j}(y_{it}) - n_{it,j}\alpha_{t,j}; n_{it,j}\mu_j\widetilde{x}_i, n_{it,j}\sigma_j)\right) \\
&\times \prod_{j \in B_{it}^1} \left(1 - \Phi(\nabla_{it,j}(y_{it}); n_{it,j}\mu_j\widetilde{x}_i, n_{it,j}\sigma_j)\right) \\
&\times \prod_{j : y_{it,j} \in \left(0, \underline{Y}_{t,j}\right], \text{ or } y_{it,j} \in (\overline{Y}_{t,j}, 1)} \phi(\nabla_{it,j}(y_{it}); n_{it,j}\mu_j\widetilde{x}_i, n_{it,j}\sigma_j) \\
&\times \prod_{j : y_{it,j} \in (\underline{Y}_{t,j}, \overline{Y}_{t,j})} \phi(\nabla_{it,j}(y_{it}) - n_{it,j}\alpha_{t,j}; n_{it,j}\mu_j\widetilde{x}_i, n_{it,j}\sigma_j) \times \left|\partial \nabla_{it,J \setminus B_{it}}\right|.
\end{aligned} \tag{10}$$

The last object in this expression, $\left|\partial \nabla_{it,J\setminus B_{it}}\right|$, is the absolute value of the Jacobian determinant of the function mapping y_{it} into $n_{it}\theta_{it}$ among indicators that do not bunch at neither 0, $\overline{Y}_{t,j}$, or 1. In our model, this is the absolute value of the determinant of the (submatrix) $(\nabla_{it} + \nabla_{it}^{\intercal})$ restricted to rows and columns corresponding to non-bunching indicators.

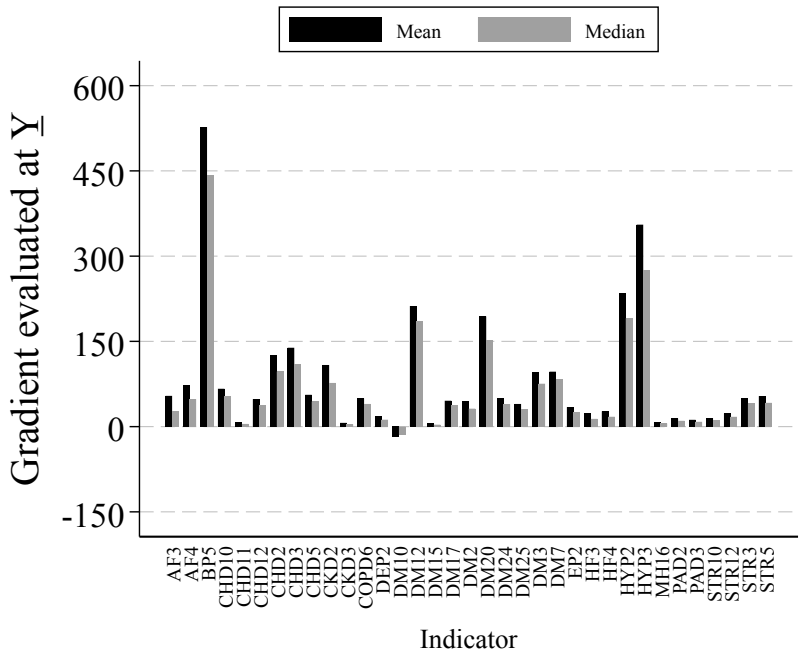
After convergence, we can use our estimates to draw θ_{it} and then evaluate the gradient of every practice payoff function at the observed $y_{it,j}$. This is centered and very concentrated around zero for all indicators, as illustrated in Appendix Figure C.5. Alternatively, we can solve the practice problem maximizing payoffs as a function of $y_{it,j}$. Doing so for every it pair we obtain a model-predicted distribution of outcomes, say $y^{\star}_{it,j}$. In Appendix Figure C.6 we show that this distribution tracks well the one observed in the data, although generally our model tends to overpredict bunching at the kink $\overline{Y}_{t,j}$ and slightly underpredict the frequency of $y_{it,j}$ exceeding $\overline{Y}_{t,j}$.

4.2 Estimation Results and Evidence of Multitasking

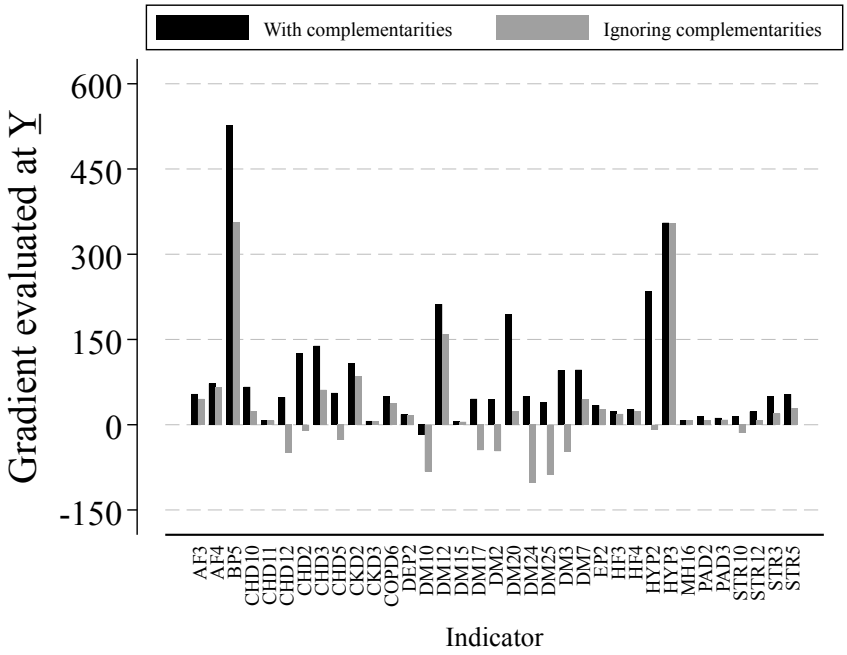
Our model contains 1060 parameters, which taken alone do not carry a lot of economic meaning. The only exception perhaps is the collection of off-diagonal terms in the matrix Λ that capture the intensity of complementarity and substitutability between indicators. In Appendix Figure C.7 we report a heatmap of these estimates. Consistent with the evidence discussed in Section 3 we find that most indicators are complements, while few are substitutes. The intensity of complementarities is highly heterogeneous. Inspecting the patterns in our estimates, it is reassuring that indicators that relate to similar medical conditions are stronger complements.

A key ingredient to study the design of physician's incentives is the gradient of the payoff function $U_{it} = u(y_{it}, n_{it}; \theta_{it}) + r_{it}$ with respect to the vector y_{it}. In the top panel of Figure 5 we display the mean and median of $\partial U_{it}/\partial y_{it,j}$ at $y_{it,j} = \underline{Y}_{t,j}$ for $j = 1, ...,36$ (after drawing the private type θ_{it} for every it pair). This shows the strength of incentives at the point $\underline{Y}_{t,j}$ at which QOF payments begin. For all but one indicators increasing the share of successes

Figure 5: First-order incentives with and without multitasking

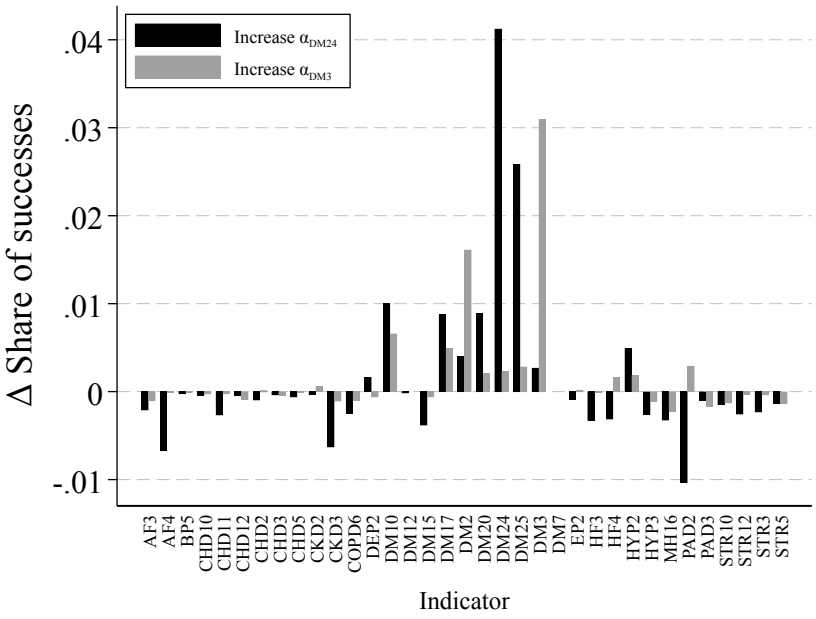


(a) Gradient of U_{it} at $y_{it} = \underline{Y}_t$



(b) Gradient of U_{it} at $y_{it} = \underline{Y}_t$; baseline vs. diagonal Λ

Figure 6: Increasing incentives by 40 GBP for one indicator



by one percentage point leads to higher payoffs. On average, for most indicators such an increase corresponds to an additional value between 30 and 150 GBP, but for some as high as 300-450.

To highlight the importance of multitasking between indicators, in the bottom panel of Figure 5 we compare baseline the gradient of payoffs evaluated at $y_{it,j} = \underline{Y}_{t,j}$ with the same quantity obtained after setting to zero all off-diagonal elements of Λ. This removes all dependencies between indicators, making the payoff of every physician additive separable. The two (average) gradients differ in meaningful ways, both in magnitude and sometimes in sign, confirming our motivating insight by which ignoring interdependencies would lead to erroneous measurement of marginal incentives. Since most indicators are complements, setting $\lambda[j,k] = 0$ whenever $j \neq k$ reduces, on average, the marginal profitability of increasing $y_{it,j}$ for all indicators.

To see how multitasking is critical to incentive design, in Figure 6, we show the simulated changes in average share of successes the 36 indicators in our analysis when we increase $\alpha_{t,\text{DM24}}$ and $\alpha_{t,\text{DM3}}$ by 40 GBP. With Λ diagonal, increasing $\alpha_{t,j}$ would only impact $y_{it,j}$. In-

stead, we find that even if the direct effects are the largest, indirect effect on other indicators can be sizable, and possibly of different sign. The patterns in Figure 6 are consistent with our estimates of Λ (Appendix Figure C.7). Sometimes, however, even if one indicator (say j) is not directly related to another (say k), if there is a third indicator ℓ for which $\lambda[j,\ell] >> 0$ while $\lambda[k,\ell] << 0$, increasing $\alpha_{t,j}$ may lead to a decrease in $y_{it,k}$.

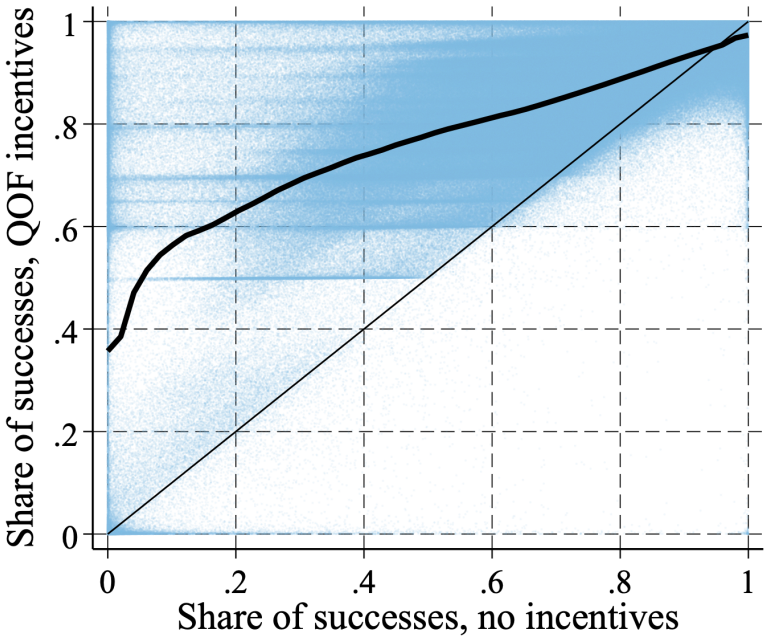
4.3 The Effect of QOF Incentives

We can now use our model to measure the overall effect of the incentives provided under the QOF. For this we set $\alpha_{t,j} = 0$ for all j, all t, and—holding fixed our draws of θ_{it} used to compute $y^{\star}_{it}$ above—compute the solution of the problem for every practice, say y^{0}_{it}. The results are illustrated in Figure 7, where we compare simulted outcomes without QOF incentives to simulated outcomes at the observed QOF incentives. We estimate that the program led, on average, to a significant increase in share of successes among the forty indicator included in our analysis.

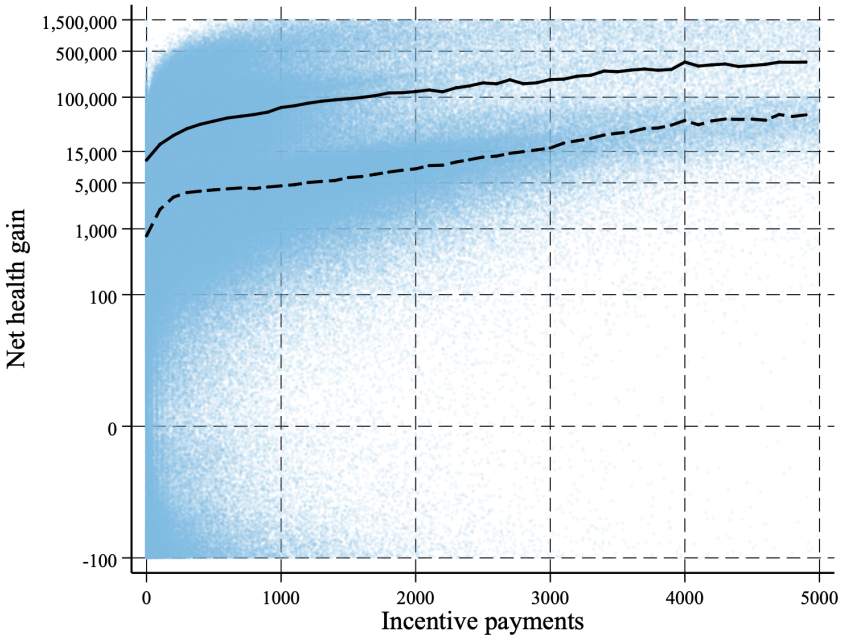
The top panel of Figure 7 compares the distribution of $y^{0}_{it,j}$ (horizontal axis) to the distribution of $y^{\star}_{it,j}$. With few exceptions due to multitasking QOF incentives shifts the distribution of outcomes upward. The solid black line indicates the conditional mean of $y^{\star}_{it,j}$ given a value of $y^{0}_{it,j}$. This shows that the QOF effect is more pronounced for practice-indicator-year triplets in which, absent incentives, success shares would have been lower than 0.6. In this region of the data we find that QOF incentives increased success shares by an average of 0.2-0.4. Most of the data, however, is concentrated in the region with higher baseline values of $y^{0}_{it,j}$, where we estimate that the increase in average outcomes generated by QOF incentives is between 0.1-0.2. Mechanically, the impact is zero (or even negative) when $y^{0}_{it,j}$ approaches 1, since the share of successes is bounded between $[0,1]$.

As described in Section 3 and Appendix B, for 18 out of the 36 indicators included in our analysis we were able to obtain measures of costs and health benefit that are used for decision making by the NHS. For this subset of indicators we can then translate our estimates

Figure 7: Impact of the QOF program



(a) Effect on share of successes



(b) Effect on health and incentive payments

of the impact of QOF incentives in easier-to-interpret, welfare-relevant quantities measured in GBP per-practice, per-year.

We do this in the bottom panel of Figure 7, where we compare the QOF payments specific to a practice-indicator-year triplet on the horizontal axis to the change (in logarithmic scale) in net QALY (i.e. increase in QALY evaluated at 25,000 GBP per-year minus cost) induced by the program for the patients in the same triplet. The solid line is the mean of changes in net QALY conditional on QOF payments, the dashed line is the median.

Also according to these measures, the effect of the program is largely positive. For practice-year pairs receiving QOF payments of 1,000 GBP for a specific indicator, the median net health gains are approximately 5,000 GBP. Since the gains in QALY are highly skewed, the average is much higher, and equal to almost 100,000 GBP. Across varying levels of incentive payments the figure suggests that the ratio between QOF payments and median net health gains is approximately 1:5.

Effects are however highly heterogeneous. Because of multitasking, QOF at times induces some practices to lower share of successes for certain indicators. This leads to the negative health gains in the bottom-left corner of the figure.

5 Empirical Design of Physician Payments

We are now ready to apply our empirical model to the problem of designing QOF incentives to maximize an objective function. Here we focus on the 18 indicators for which we were able to collect health benefits and costs as measured by the NHS. In what follows, we use ω_j^h to denote the increase in health (with one QALY being valued at 25,000 GBP per-person per-year) when indicator j is a success for a given patient, and ω_j^c to denote the increase in medical costs. Moreover, to limit computational burden we hold fixed $\underline{Y}_t$, which is rarely binding in the data, and set $\overline{Y}_{t,j}$ equal to 1 for all indicators. Therefore our design problem only optimizes over α. These simplifications imply that the improvements we generate represents a lower bound to the solution of a more flexible design problem.

The problem can be written as

$$W^{\star} = \max_{\alpha} \sum_{j} \sum_{i} \sum_{t} n_{it,j} \int y_{it,j} \left(\omega_j^h - \omega_j^c\right) - \rho(y_{it,j}; \alpha_j) dF\left(\theta_{it} \middle| x_i, \widehat{\xi}_i\right) \tag{11}$$
$$s.t.\ y_{it} = \arg\max_{y} \sum_{j} n_{it,j} \left(\rho(y_j; \alpha_j) + y_j \left(\theta_{i,j} - \left(\sum_{k} y_k \times \lambda[j,k]\right)\right)\right).$$

Notice that solving this problem requires to solve, for every value of α, one separate optimization problem for every i, t, and—theoretically—θ_{it} combination. To the best of our understanding, this is unfeasible.

Therefore, we approximate the problem in two ways. First, we (k-means) cluster practice-year pairs in G (currently $G = 20$) groups with similarities in n_{it}, x_i, and $\widehat{\xi}_i$. For every such group $g = 1, 2, ..., G$ we compute the average of each of these observables, say n_g, x_g, and $\widehat{\xi}_g$. Let π_g be the share of it pairs in each group. Then, for every group we approximate the integral in (11) by taking R (currently $R = 100$) draws of the private type (now θ_g, rather than θ_{it}) conditional on $(x_g, \widehat{\xi}_g)$. Finally, we can feasibly solve

$$\widehat{W} = \max_{\alpha} \frac{1}{R} \sum_{j} \sum_{g} \pi_g n_{g,j} \sum_{r} y_{g,j}^r \left(\omega_j^h - \omega_j^c\right) - \rho(y_{g,j}; \alpha_j) \tag{12}$$
$$s.t.\ y_g^r = \arg\max_{y} \sum_{j} n_{g,j} \left(\rho(y_j; \alpha_j) + y_j \left(\theta_{g,j}^r - \left(\sum_{k} y_k \times \lambda[j,k]\right)\right)\right).$$

Of course, $\widehat{W} \leq W^{\star}$, so that our results will again be a lower bound to the potential impact of designing QOF incentives optimally.

Once we obtain the solution $\widehat{\alpha}$ from (12), we apply it to the entire distribution of it pairs, drawing $\theta_i t$. This allows us to compare the resulting outcomes, say $\widehat{y}_{it}$, to the QOF outcomes $y_{it}^{\star}$ and outcomes without incentives y_{it}^0. For all three simulations we focus only on the 18 indicators for which we observe QALY and costs, and hold the draws of θ_{it} fixed.

Table 3 summarizes our (preliminary) findings. We estimate that removing QOF incentives would lower welfare by 5%: the reduction in payments (353 million) and medical costs

Table 3: Counterfactual design of QOF incentives

	No QOF Δ from QOF	QOF	Optimized QOF Δ from QOF
Practice payoffs	-348 -11%	3,240	164 5%
QOF payments	-353 -100%	353	199 56%
Medical costs	-1,431 -3%	43,189	683 2%
Health benefits	-5,553 -4%	131,565	3,857 3%
Welfare	-4,117 -5%	91,264	3,139 3%

Notes: All monetary values are in GBP millions. Welfare is computed as the Practice payoffs + Health benefits - QOF payments - Medical costs

(1.4 billion) would not compensate the reduction in QALY's (5.6 billion). On the other hand, there seems to be room for significant improvement in incentive design. Table 3 suggests that re-optimizing incentives would—despite increasing QOF payments by 199 million and costs by 683 million—lead to additional health benefits worth almost 4 billion (+3% from baseline). Physicians' payoffs would also increase by approximately 5%, adding 164 million to the baseline level of 3.2 billion GBP. These changes would add up to a 3% increase in the objective defined in (11); this does not currently include physician's payoffs.

6 Conclusions

We proposed an empirically tractable principal-agent model that includes multitasking and private information. The model is identified when assuming linear-quadratic utility, and when it is possible to observe rich variation in incentives combined with exogenous variation in exposure to incentives across agents. In practice, this requires instrumental variables that affect the composition of tasks across agents but are independent from payoff-relevant

private information.

We apply our method to the pay-for-performance program for primary care physicians in England during the 2009-2019 period (Quality of Outcomes Framework; QOF). We find ample evidence for physicians being responsive to incentives and for dependencies between different clinical outcomes as measured by revenue-relevant QOF indicators. Exploiting variation in the location of physicians' practices as exogenous, we estimate our model, and detect meaningful (and mostly complementary) interactions between indicators.

Counterfactual simulations show that the program led to significant improvements in health that—valuing health using National Health Service criteria—largely outweigh payments to physicians. Optimizing the design of incentive may improve outcomes further.

References

Berry, Steve, Ahmed Khwaja, Vineet Kumar, Andres Musalem, Kenneth C Wilbur, Greg Allenby, Bharat Anand, Pradeep Chintagunta, W Michael Hanemann, Przemek Jeziorski et al., "Structural models of complementary choices," *Marketing Letters*, 2014, *25*, 245–256.

Berry, Steven T and Philip A Haile, "Identification in differentiated products markets using market level data," *Econometrica*, 2014, *82* (5), 1749–1797.

Brown, Zach Y, Christopher Hansman, Jordan Keener, and Andre F Veiga, "Information and Disparities in Health Care Quality: Evidence from GP Choice in England," Technical Report, National Bureau of Economic Research 2023.

Buser, Thomas and Noemi Peter, "Multitasking," *Experimental economics*, 2012, *15* (4), 641–655.

Chalkley, Martin and James M Malcomson, "Government purchasing of health services," *Handbook of health economics*, 2000, *1*, 847–890.

Choné, Philippe and Ching to Albert Ma, "Optimal health care contract under physician agency," *Annals of Economics and Statistics/Annales d'Économie et de Statistique*, 2011, pp. 229–256.

Clemens, Jeffrey and Joshua D Gottlieb, "Do physicians' financial incentives affect medical treatment and patient health?," *American Economic Review*, 2014, *104* (4), 1320–1349.

Dinerstein, Michael and Isaac M Opper, "Screening with Multitasking: Theory and Empirical Evidence from Teacher Tenure Reform," Technical Report, National Bureau of Economic Research 2022.

Dumont, Etienne, Bernard Fortin, Nicolas Jacquemet, and Bruce Shearer, "Physicians' multitasking and incentives: Empirical evidence from a natural experiment," *Journal of health economics*, 2008, *27* (6), 1436–1450.

Dunn, Abe, Joshua D Gottlieb, Adam Hale Shapiro, Daniel J Sonnenstuhl, and Pietro Tebaldi, "A denial a day keeps the doctor away," *The Quarterly Journal of Economics*, 2024, *139* (1), 187–233.

Einav, Liran, Amy Finkelstein, and Neale Mahoney, "Provider incentives and healthcare costs: Evidence from long-term care hospitals," *Econometrica*, 2018, *86* (6), 2161–2219.

_ , _ , Yunan Ji, and Neale Mahoney, "Voluntary regulation: Evidence from Medicare payment reform," *The Quarterly Journal of Economics*, 2022, *137* (1), 565–618.

Gaynor, Martin, James B Rebitzer, and Lowell J Taylor, "Physician incentives in health maintenance organizations," *Journal of Political Economy*, 2004, *112* (4), 915–931.

_ , Nirav Mehta, and Seth Richards-Shubik, "Optimal Contracting with Altruistic Agents: Medicare Payments for Dialysis Drugs," *American Economic Review*, 2023, *113* (6), 1530–1571.

Gillam, Stephen J, A Niroshan Siriwardena, and Nicholas Steel, "Pay-for-performance in the United Kingdom: impact of the quality and outcomes framework—a systematic review," *The Annals of Family Medicine*, 2012, *10* (5), 461–468.

Godager, Geir and Daniel Wiesen, "Profit or patients' health benefit? Exploring the heterogeneity in physician altruism," *Journal of health economics*, 2013, *32* (6), 1105–1116.

Goes, Paulo B, Noyan Ilk, Mingfeng Lin, and J Leon Zhao, "When more is less: Field evidence on unintended consequences of multitasking," *Management Science*, 2018, *64* (7), 3033–3054.

Gupta, Atul, "Impacts of performance pay for hospitals: The readmissions reduction program," *American Economic Review*, 2021, *111* (4), 1241–1283.

Hannaway, Jane, "Higher order skills, job design, and incentives: An analysis and proposal," *American educational research journal*, 1992, *29* (1), 3–21.

Holmström, Bengt, "Moral hazard and observability," *The Bell journal of economics*, 1979, pp. 74–91.

Holmstrom, Bengt and Paul Milgrom, "Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design," *The Journal of Law, Economics, and Organization*, 1991, *7* (special_issue), 24–52.

Hong, Fuhai, Tanjim Hossain, John A List, and Migiwa Tanaka, "Testing the theory of multitasking: Evidence from a natural field experiment in Chinese factories," *International Economic Review*, 2018, *59* (2), 511–536.

Kim, Minkyung, K Sudhir, and Kosuke Uetake, "A structural model of a multitasking salesforce: Incentives, private information, and job design," *Management Science*, 2022, *68* (6), 4602–4630.

Lewbel, Arthur and Lars Nesheim, "Sparse demand systems: corners and complements," Technical Report, cemmap working paper 2019.

Li, Jinhu, Jeremiah Hurley, Philip DeCicca, and Gioia Buckley, "Physician response to pay-for-performance: Evidence from a natural experiment," *Health economics*, 2014, *23* (8), 962–978.

Manthei, Kathrin and Dirk Sliwka, "Multitasking and subjective performance evaluations: Theory and evidence from a field experiment in a bank," *Management Science*, 2019, *65* (12), 5861–5883.

McGuire, Thomas G, "Physician agency," *Handbook of health economics*, 2000, *1*, 461–536.

Mirrlees, James A, "The optimal structure of incentives and authority within an organization," *The Bell Journal of Economics*, 1976, pp. 105–131.

Mullen, Kathleen J, Richard G Frank, and Meredith B Rosenthal, "Can you get what you pay for? Pay-for-performance and the quality of healthcare providers," *The Rand journal of economics*, 2010, *41* (1), 64–91.

Rodríguez-Lesmes, Paul and Marcos Vera-Hernández, "Multitasking, Two-part Contracts, and Bunching: an Application to Doctors' Jobs and Incentive Contracts," 2021. *Mimeo, University College London.*

Ross, Stephen A, "The economic theory of agency: The principal's problem," *The American economic review*, 1973, *63* (2), 134–139.

Shi, Maggie, "Monitoring for waste: Evidence from medicare audits," *The Quarterly Journal of Economics*, 2024, *Forthcoming.*

Slade, Margaret E, "Multitask agency and contract choice: An empirical exploration," *International Economic Review*, 1996, pp. 465–486.

Wilson, R, "The Structure of Incentives for Decentralization under Uncertainty, La Decision, Editions du Centre National de la Rocherche Scientifique.-54," 1969.

Appendix

A Proof of Proposition 1

To begin, under Assumption 2 we can invoke Berry and Haile (2014) and treat $\xi_{it} = \widehat{\xi}_{it}$ is known hereafter. Therefore, in what follows we condition on $(x_{it}, \widehat{\xi}_{it})$ throughout but omit this to keep notation uncluttered without any loss of generality; by Assumption 2, then, in what follows $f_j(\theta_{it,j}|n_{it}) = f_j(\theta_{it,j})$.

Next, for all it, and all j whenever $y_{it,j} \in (0,1)$ the following FOC must hold:

$$\theta_{it,j} = -\rho_j'(y_{it,j}) + \sum_k w_{it}(j,k) \times \lambda[j,k] \times y_{it,k}, \quad \text{where } w_{it}(j,k) = \frac{n_{it,j} + n_{it,k}}{n_{it,j}}, \tag{13}$$

and independence in Assumption 2 implies that the observed density $g(\cdot)$ of outcomes is

$$g(y_{it}|n_{it}) = \prod_j f_j \left(\underbrace{-\rho_j'(y_{it,j}) + \sum_k w_{it}(j,k) \times \lambda[j,k] \times y_{it,k}}_{\theta_{it,j} \text{ in (13)}} \right) \times \left| \partial \widetilde{\Lambda}_{it} \right|, \tag{14}$$

where $\left|\partial \widetilde{\Lambda}_{it}\right|$ is the absolute value of the determinant of the Jacobian of the function mapping y_{it} into θ_{it}. Here, this is the determinant of the matrix with generic row j and column k equal to $w_{it}(j,k)\lambda[j,k]$. Note that, while Λ is symmetric, $\left|\partial \widetilde{\Lambda}_{it}\right|$ is not, since it is generally the case that $w_{it}(j,k) \neq w_{it}(k,j)$.

The left-hand side is observed, and so are its derivatives. Then, whenever $g(y_{it}|n_{it}) > 0$, we can calculate for every j

$$\begin{aligned} \partial \log(g(y_{it}|n_{it}))/\partial y_{it,j} = &\frac{f_j'(\theta_{it,j})}{f_j(\theta_{it,j})} \left(-\rho_j''(y_{it,j}) + w_{it}(j,j) \times \lambda[j,j] \right) \\ &+ \sum_{k \neq j} \frac{f_k'(\theta_{i,k})}{f_k(\theta_{i,k})} \left(w_{it}(j,k) \times \lambda[j,k] \right). \end{aligned} \tag{15}$$

Even if f_j were known, the equations in (15)—one for every j—represent an "underidentified" system since Λ contains $J(J-1)/2$ distinct parameters.

Differentiating (15) with respect to $n_{it,k}$, $k \neq j$ we obtain

$$\begin{aligned}
&\partial^2 \log(g(y_{it}|n_{it}))/\partial y_{it,j}\partial n_{it,k} = \\
&\partial\left(\frac{f_j'(\theta_{it,j})}{f_j(\theta_{it,j})}\right)/\partial n_{it,k}\left(-\rho_j''(y_{it,j}) + w_{it}(j,j)\times\lambda[j,j]\right) + \frac{f_k'(\theta_{it,k})}{f_k(\theta_{it,k})}\left(\frac{\lambda[j,k]}{n_{it,j}}\right) \\
&+\sum_{\ell\neq j}\partial\left(\frac{f_\ell'(\theta_{it,\ell})}{f_\ell(\theta_{it,\ell})}\right)/\partial n_{it,k}\left(w_{it}(j,\ell)\times\lambda[j,\ell]\right).
\end{aligned} \tag{16}$$

Together, (15) and (16) define a system of $J + J(J-1)$ independent equations. Therefore, if the terms

$$\frac{f_j'(\theta_{it,j})}{f_j(\theta_{it,j})},\ \partial\left(\frac{f_j'(\theta_{it,j})}{f_j(\theta_{it,j})}\right)/\partial n_{it,k},\ \ j\neq k \tag{17}$$

were known, the completeness condition in Assumption 3 would be sufficient to identify Λ.[8]

The first part of Assumption 3 implies that the following expressions are well defined for all j, all k:

$$\partial \log(g(y_{it}|n_{it}))/\partial\alpha_{t,j} = \frac{f_j'(\theta_{it,j})}{f_j(\theta_{it,j})}\left(-\partial^2\rho(y_{it,j};\alpha_{t,j})/\partial y_{it,j}\partial\alpha_{t,j}\right); \tag{18}$$

$$\partial^2 \log(g(y_{it}|n_{it}))/\partial\alpha_{t,j}\partial n_{it,k} = -\left(\partial^2\rho(y_{it,j};\alpha_{t,j})/\partial y_{it,j}\partial\alpha_{t,j}\right)\times\partial\left(\frac{f_j'(\theta_{it,j})}{f_j(\theta_{it,j})}\right)/\partial n_{it,k}. \tag{19}$$

Since the left-hand side is observed, and ρ is a known primitive, (18) and (19) identify the

[8]Indeed, if two distinct Λ, $\widehat{\Lambda}$ were such that, for both, taking expectations over combinations of i,t,j,k would satisfy (15) and (16) in expectation almost surely, then this would imply that (almost surely) for all j and all k

$$0 = \frac{f_j'(\theta_{it,j})}{f_j(\theta_{it,j})}\left(-\rho_j''(y_{it,j}) + w_{it}(j,j)\times\left(\lambda[j,j]-\widehat{\lambda}[j,j]\right)\right) + \sum_{k\neq j}\frac{f_k'(\theta_{i,k})}{f_k(\theta_{i,k})}\left(w_{it}(j,k)\times\left(\lambda[j,k]-\widehat{\lambda}[j,k]\right)\right);$$

$$\begin{aligned}
0 = \partial\left(\frac{f_j'(\theta_{it,j})}{f_j(\theta_{it,j})}\right)/\partial n_{it,k}\left(-\rho_j''(y_{it,j}) + w_{it}(j,j)\times\left(\lambda[j,j]-\widehat{\lambda}[j,j]\right)\right) + \frac{f_k'(\theta_{it,k})}{f_k(\theta_{it,k})}\left(\frac{\left(\lambda[j,k]-\widehat{\lambda}[j,k]\right)}{n_{it,j}}\right) \\
+\sum_{\ell\neq j}\partial\left(\frac{f_\ell'(\theta_{it,\ell})}{f_\ell(\theta_{it,\ell})}\right)/\partial n_{it,k}\left(w_{it}(j,\ell)\times\left(\lambda[j,\ell]-\widehat{\lambda}[j,\ell]\right)\right);
\end{aligned}$$

But since this is an overidentified system of equations these conditions can only be true if $\Lambda = \widehat{\Lambda}$.

terms in (17). Therefore, Λ is identified.

Lastly, equation (13) can be used to identify nonparametrically the family of distributions $f_j(\cdot)$, one for every j. This completes the proof.■

B Details of Data Sources

The QOF consists of a set of achievement measures (known as indicators) that cover various aspects of patient care. Practices score points based on their level of achievement in each indicator. Points are then converted into payments according to a predetermined formula that also considers practice size and disease prevalence relative to national average values. The NHS determines the total number of points in the QOF and their allocation, as well as the monetary value of each point, on an annual basis.[9]

We obtained public QOF data for the entire period of analysis (2009-2019) from NHS Digital and from the UK Government Asset Portal. We collect data on indicators and their assigned points, disease prevalence, practice achievement, and the monetary value of QOF points. We select 40 indicators from the QOF Clinical Domain, excluding those indicators that are of "Register" category (e.g., "The practice can produce a register of patients with atrial fibrillation"). We then remove four of these 40 indicators for which the average number of patients in the practice year is fewer than 10 patients. These data allow us to construct the terms that enter equation (5) in Section 3.

For each practice i, year t, and indicator j, we observe the share of successes ($y_{it,j}$), and can calculate the "floor" ($\underline{Y}_{t,j}$) and the "ceiling" ($\overline{Y}_{t,j}$).[10] Finally, we calculate the monetary amount received per-relevant-patient for every percentage point increase in the

[9] See also https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2022-23/technical-annex, accessed March 28, 2024.

[10] The denominator used to compute the share of successes is the number of patients in the practice-year that are relevant for a specific indicator. This number does not include exceptions.

share of successes up until the ceiling as:

$$\alpha_{t,j} = \frac{M_{t,j} \times V_t \times \text{CPI}_t}{(\overline{Y}_{t,j} - \underline{Y}_{t,j}) \times \mathbb{E}[T_t] \times \mathbb{E}\left[\frac{n_{t,g(j)}}{T_t}\right]}.$$

In this expression, $M_{t,j}$ represents the maximum QOF points practices can earn for indicator j, V_t is the monetary value of a QOF point, CPI_t is the consumer price index,[11] $\mathbb{E}[T_t]$ represents the average number of patients at each practice and $\mathbb{E}\left[\frac{n_{t,g(j)}}{T_t}\right]$ represents the average share of patients in the indicator group to which indicator j belongs. We complement these data with the costs and benefits (measured in Quality-Adjusted Life Year, QALY) associated with 18 QOF indicators, obtained from the National Institute for Health and Care Excellence and from Department of Health reports. We convert QALYs to GBP using the societal value of a QALY used by the National Institute for Health and Care Excellence, which is £25,000.

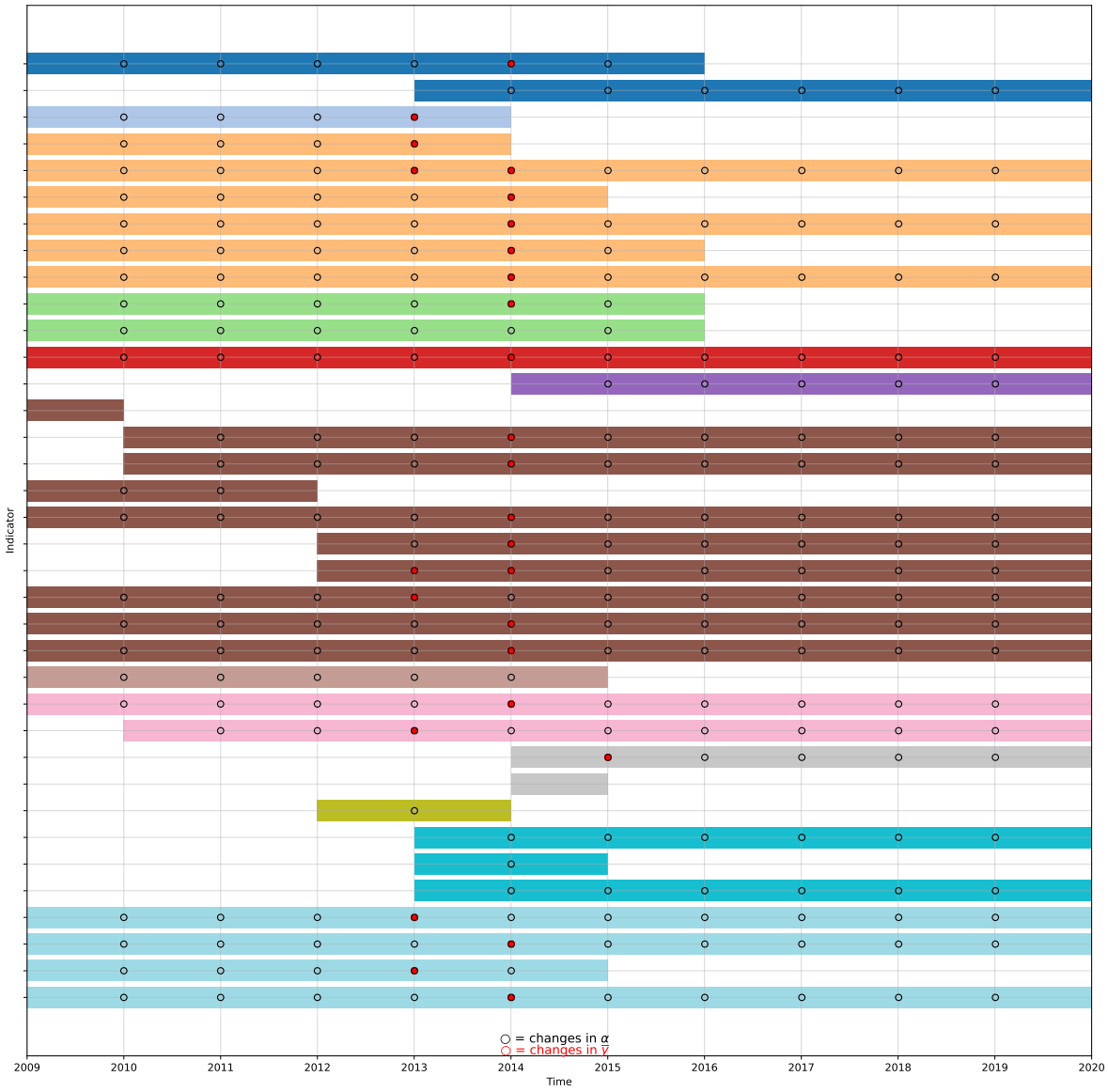
Besides QOF data, we gather additional information on GP practices in England. First, we obtain practice-level data on the number of GPs, their average age, and their qualification. These characteristics are available only for the 2015-2019 period, but since our main data covers the whole 2009-2019 period, we extend the data to the earlier years by taking the first available value for each practice. Second, we obtain yearly data (for 2015-2019) on the number of patients within each Lower Layer Super Output Areas (LSOA) registered with each GP practice.[12] We obtained both sets of data from NHS Digital. Finally, we collect data on the locations of GP practices and LSOAs from NHS Digital and the Office for National Statistics, respectively.

[11] We download annual inflation rates computed by the Office for National Statistics from Statista.

[12] An LSOA is a geographic area used for statistical reporting in England and Wales. There are approximately 35,000 LSOAs in England, each containing an average of around 1,500 residents or 650 households.

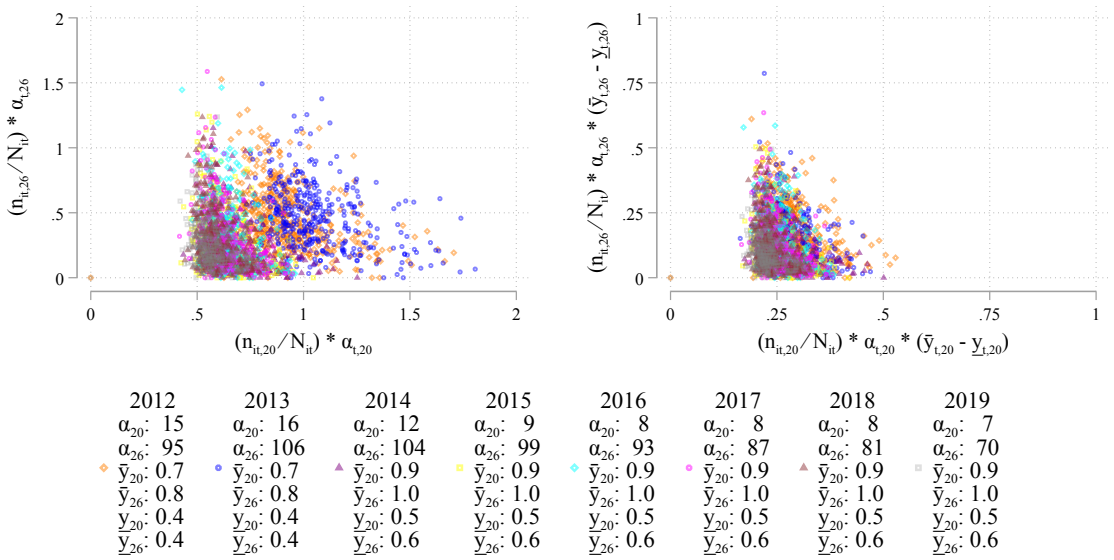
C Additional Tables and Figures

Figure C.1: QOF indicator timelines



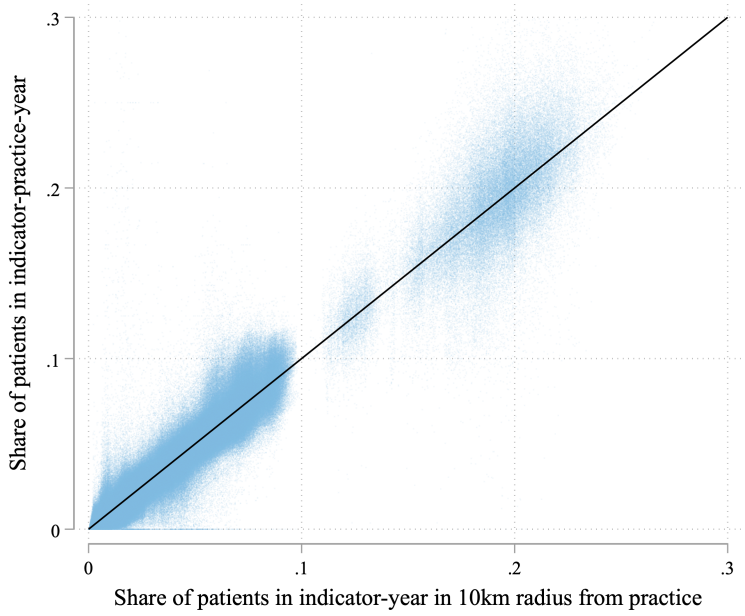
Notes: Timelines for which each of the indicators is included in the QOF. Indicators are colored according to the indicator group to which they belong.

Figure C.2: Identifying variation in payment schedule and number of patients

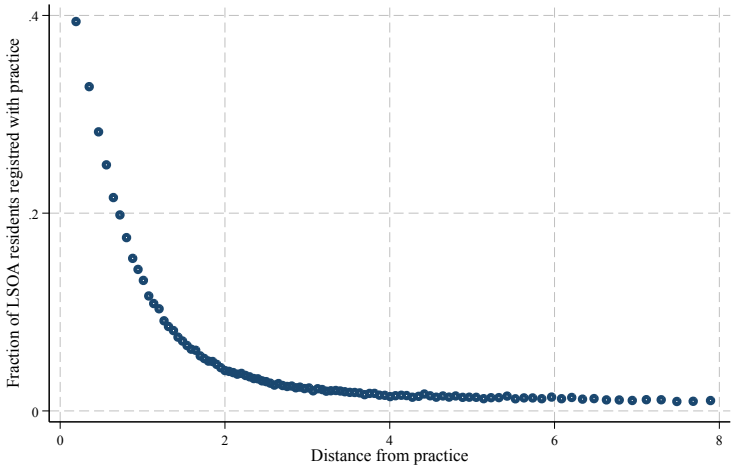


Notes: Scatterplots for a subsample of 5% of the data for indicators 20 (DM2) and 26 (HF3). Indicator 20 rewards practices for maintaining blood pressure below a relevant threshold in diabetic patients, while indicator 26 rewards practices for treating patients with heart failure with heart-protective medications (ACE inhibitors or Angiotensin Receptor Blockers).

Figure C.3: Demand estimation intuitions

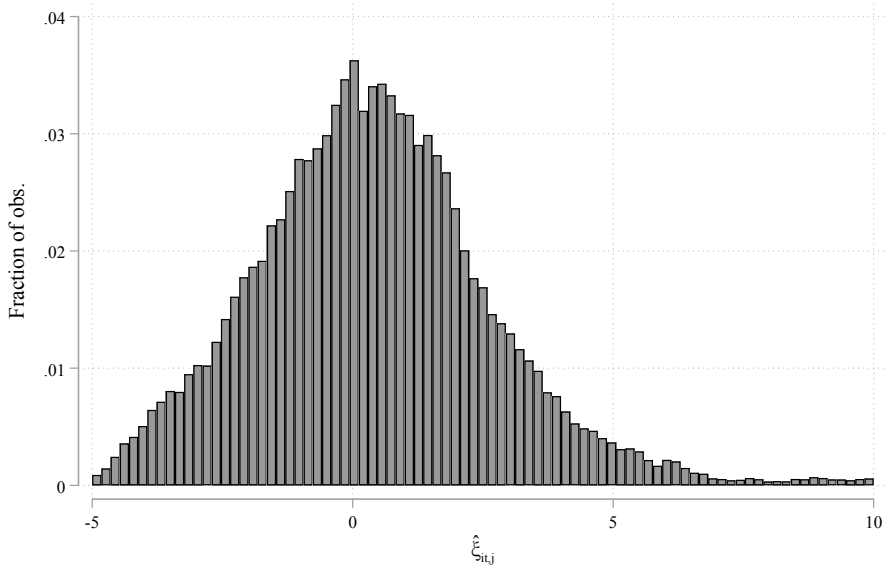


(a) Patient share vs patient share in practices within 10km radius (with 45-degree line)

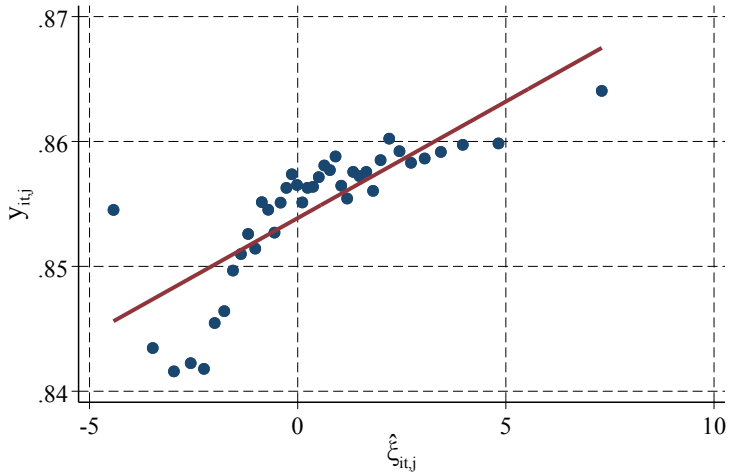


(b) Fraction of LSOA residents registered with a practice vs distance of that practice from LSOA centroid (in km)

Figure C.4: Estimated demand unobservables



(a) Distribution of estimated demand unobservables



(b) Binned scatterplot of share of successes and estimated demand unobservables

Figure C.5: Gradient at data

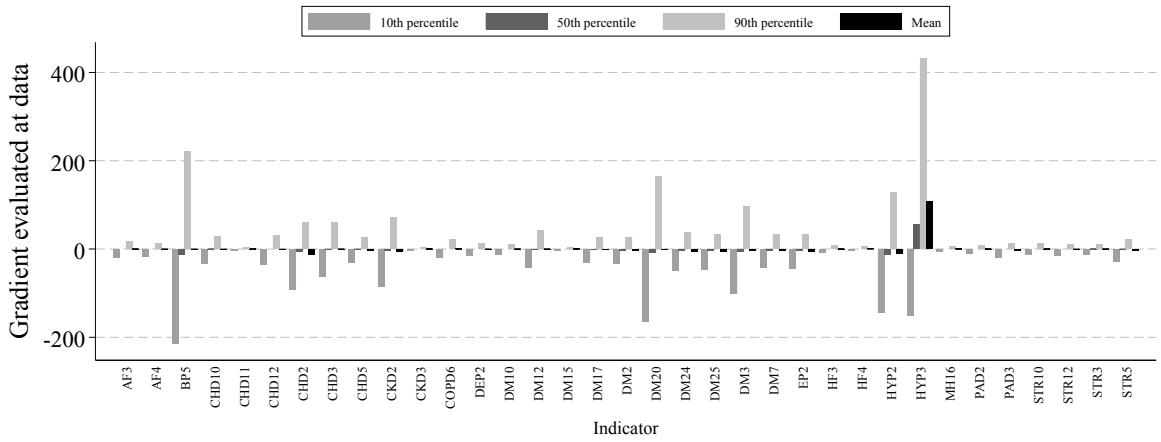


Figure C.6: Goodness of fit

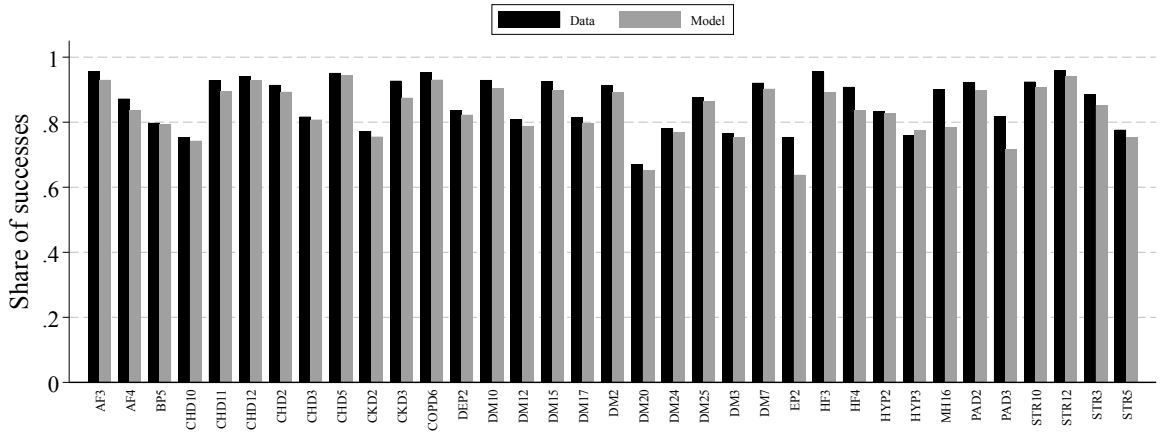
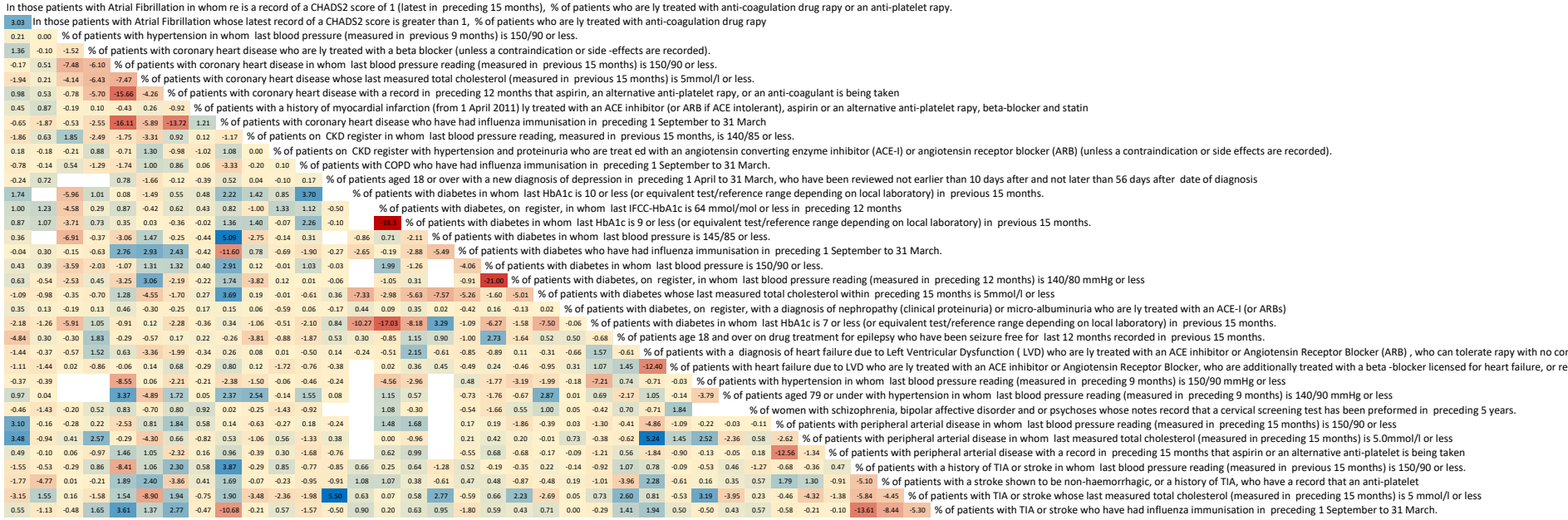


Figure C.7: Estimates of off-digaonal elements of Λ



D Robustness of demand estimates to different radii

A crucial component of our demand estimation is the distance between practices and LSOA. We often select practices or LSOAs close enough to each other by selecting only the practices or LSOAs within a given radius of each other. We use two different radii in our demand estimation: r_1 is used to estimate the composition of patients within LSOAs, while r_2 is used to define markets.

1. We use r_1 to estimate the number of patients in LSOA ℓ, year t, and indicator j. Specifically, we do the following:

   (a) We sum the number of patients in each indicator j across practices k within a radius r_1 of each practice i.

   (b) We then estimate the fraction of LSOA ℓ's patients in indicator j as the ratio between

      i. The sum of the number of patients around practice i in indicator j across practices within a radius r_1 of LSOA ℓ; and

      ii. The sum of the number of patients across LSOAs ℓ' within a radius r_1 of ℓ.

   (c) Finally, we estimate the number of patients in LSOA ℓ, year t, and indicator j as the product of the number of patients in ℓ and the estimated fraction of ℓ's patients in indicator j (computed in the previous point).

2. We use r_2 to define markets in the demand estimation. Specifically, we do the following:

   (a) We compute the market share of practice i in year t and indicator j as the ratio between

      i. The number of patients in practice i, year t, indicator j from the QOF data; and

      ii. The sum of the estimated number of patients in indicator j across LSOAs ℓ (computed in point 1.(c) above) that are within a radius r_2 of i.

(b) We then construct a matrix storing the distribution of patients across LSOAs within a radius r_2 of practice i. The entries of this matrix are the elements $\psi_{t,j}(i,\ell)$, which we use to weigh the contribution of each LSOA in computing the model-predicted market shares for each practice. These contributions are zero whenever the distance between the LSOA and the practice is greater than r_2.

(c) Finally, to compute the model-predicted market share, we perform a summation over LSOAs within a radius r_2 of practice i. The contribution of each LSOA is weighted by the matrix construted in the previous point. Each factor in the summation over LSOAs has at the denominator a summation over all practices i' within a radius r_2 of the given LSOA.

In our baseline demand estimation, we set $r_1 = r_2 = 10$ km.

In this exercise, we repeat our demand estimation using 16 different pairs of radii (r_1, r_2) from a grid $(r_1, r_2) \in \{4, 8, 12, 16\} \times \{4, 8, 12, 16\}$.[13] We then assess how the results of the demand estimation, based on a specific pair of radii, compare with our baseline. Specifically, we consider four different outcomes:

1. The correlation between the residuals obtained with a given pair of radii (r_1, r_2) and the residuals obtained in our baseline:

$$\text{corr}(\hat{\xi}(r_1, r_2), \hat{\xi}(10, 10))$$

this outcome is plotted in the top left panel of Figure D.8.

2. The standard deviation of the residuals obtained with a given pair of radii (r_1, r_2):

$$\text{SD}(\hat{\xi}(r_1, r_2))$$

this outcome is plotted in the top right panel of Figure D.8.

[13] All radii are measured in kilometers.

3. The standard deviation of the mean utility $\delta^*_{it,j}$ obtained with a given pair of radii (r_1, r_2):

$$\text{SD}(\delta^*_{it,j}(r_1, r_2))$$

this outcome is plotted in the bottom left panel of Figure D.8.

4. The average estimated distance coefficient α^*_j (in absolute value) obtained with a given pair of radii (r_1, r_2):

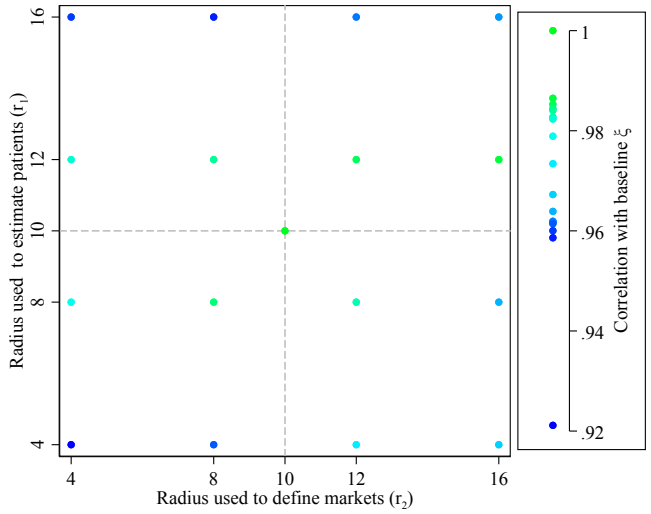
$$\text{AVERAGE}(\alpha^*_j(r_1, r_2))$$

this outcome is plotted in the bottom right panel of Figure D.8.

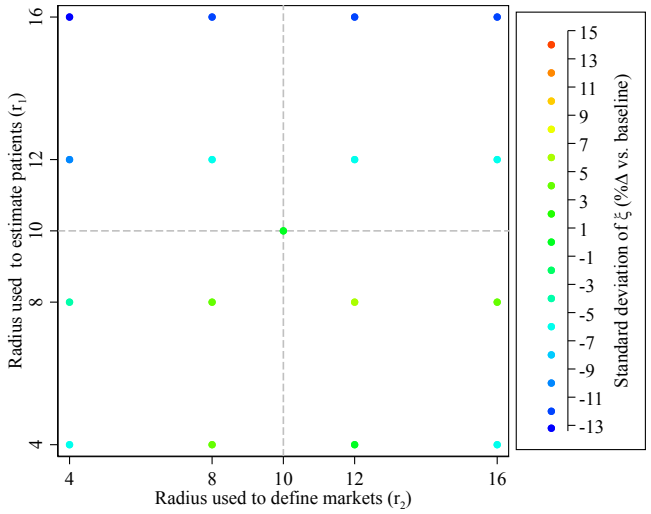
For measures 2-4, we consider the percentage change with respect to the corresponding outcomes from our baseline demand estimation.

There are three main takeaways from this robustness exercise. First, the residuals from all pairs of radii show a high correlation with the baseline, indicating the robustness of our results. The smallest correlation we detect is equal to 0.92. Second, the changes in the standard deviation of residuals and mean utility are generally fairly small, showing decreases that are smaller than 15%. Last, the distance coefficient shows some sensitivity to the choice of radii. Specifically, smaller radii used in the market definition (r_2) are associated with smaller average distance coefficients (in absolute value).

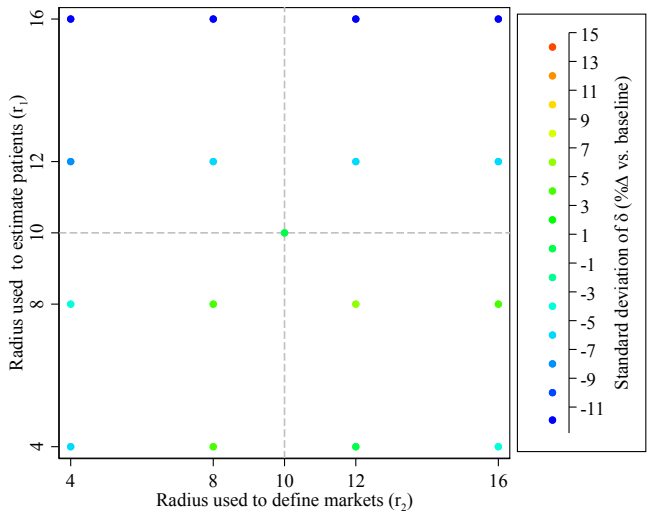
Figure D.8: Demand robustness to alternative radii



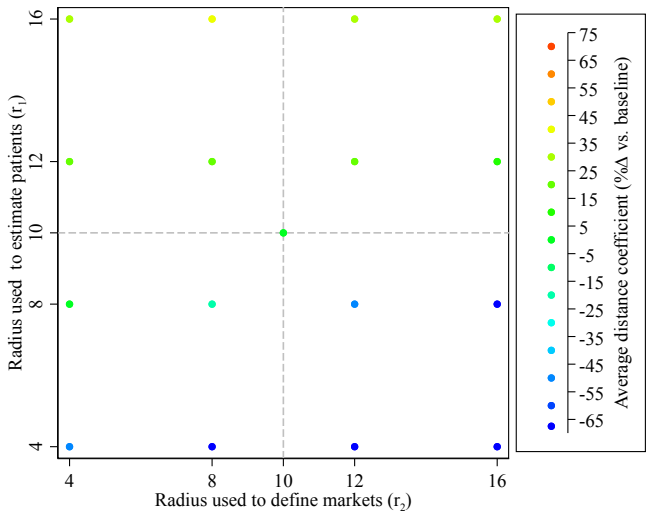
(a) $\text{corr}(\hat{\xi}(r_1, r_2), \hat{\xi}(10, 10))$



(b) $\text{SD}(\hat{\xi}(r_1, r_2))$



(c) $\text{SD}(\delta^*_{it,j}(r_1, r_2))$



(d) $\text{AVERAGE}(\alpha^*_j(r_1, r_2))$

Notes: *All four panels always display the baseline value in the center of the grid in green. Each dot in panel (a) shows the correlation between the residuals obtained with a given pair of radii* (r_1, r_2) *and the residuals obtained in our baseline: bluer dots signify lower correlations. Panels (b)-(d) show the standard deviation of* $\hat{\xi}_{it,j}$*, the standard deviation of* $\delta^*_{it,j}$*, and the average estimated distance coefficient (in absolute value), respectively. All three measures are in percentage changes with respect to the baseline. Values close to 0 are in shades of green, percent decreases are in shades of blue, and percent increases are in shades of red..*

E Illustrative Framework

There are J tasks (or output goods). There is an input vector, $t \in \mathbb{R}_+^J$ with component $t_j \in \mathbb{R}_+$. For instance, t_j is the amount of time dedicated to task j. Let ϕ_j be the unit cost of input t_j. Then, total cost is

$$C = \sum_j \phi_j t_j$$

We will consider production functions of the form $f(t) : \mathbb{R}_+^J \to [0,1]^J$ with component $f_j(t) : \mathbb{R}_+^J \to [0,1]$. Let $f(t)$ be a $J \times 1$ column vector. That is, $f_j(t)$ is the resulting output for task j when the vector of inputs is t.

Proposition 2. *For any positive definite matrix Λ, there exists a production function $f(t)$ such that cost can be written as a quadratic form of the outputs, i.e.,*

$$C = \sum_j \phi_j t_j = f(t)^T \Lambda (f(t)).$$

Proof. If Λ is positive definite, we can write $\Lambda = DZD^T$ where D is orthonormal and Z is diagonal and contains the eigenvalues of Λ, which are real and positive. Since D is orthonormal, $D^{-1} = D^T$. Let $f(t) = Dg(t)$, for some function $g(t)$ to be specified below. Then cost is

$$\begin{aligned} C &= (Dg)^T DZD^T Dg \\ &= gD^{-1}DZD^{-1}Dg \\ &= gZg \end{aligned}$$

Let χ_j be the j-th eigenvalue of Λ (i.e., the j-th entry of the diagonal matrix Z). Let g have

components $g_j(t) = \sqrt{\frac{\phi_j t_j}{\chi_j}}$. Since Z is diagonal, we obtain

$$C = \sum_j \left(\sqrt{\frac{\phi_j t_j}{\chi_j}} \right)^2 \chi_j = \sum_j \phi_j t_j.$$

□

If the outputs are independent, then $\Lambda = Z$ is diagonal, D is the identity matrix and $g_j(t) = \sqrt{\phi_j t_j}$. In our application, the output is $y = f(t)$ and $\phi_j = 1, \forall j$.