http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 67, No. 4, July–August 2019, pp. 1002–1026
ISSN 0030-364X (print), ISSN 1526-5463 (online)

CONTEXTUAL AREAS

Data-Driven Incentive Design in the Medicare Shared Savings Program

Anil Aswani,[a] Zuo-Jun Max Shen,[a,b,c] Auyon Siddiq[d]

[a] Department of Industrial Engineering and Operations Research, University of California, Berkeley, California 94720; [b] Department of Civil and Environmental Engineering, University of California, Berkeley, California 94720; [c] Tsinghua-Berkeley Shenzhen Institute, University of California, Berkeley, California 94720; [d] Anderson School of Management, University of California, Los Angeles, California 90095

Contact: aaswani@berkeley.edu (AA); maxshen@berkeley.edu, http://orcid.org/0000-0003-4538-8312 (Z-JMS); auyon.siddiq@anderson.ucla.edu, http://orcid.org/0000-0003-2977-5558 (AS)

Received: November 21, 2016
Revised: December 4, 2017; June 10, 2018
Accepted: September 5, 2018
Published Online in Articles in Advance: May 28, 2019

Subject Classifications: hospitals: healthcare; data analysis: statistics; applications: integer programming
Area of Review: Operations and Supply Chains

https://doi.org/10.1287/opre.2018.1821

Abstract. The Medicare Shared Savings Program (MSSP) was created under the Patient Protection and Affordable Care Act to control escalating Medicare spending by incentivizing providers to deliver healthcare more efficiently. Medicare providers that enroll in the MSSP earn bonus payments for reducing spending to below a risk-adjusted financial benchmark that depends on the provider's historical spending. To generate savings, a provider must invest to improve efficiency, which is a cost that is absorbed entirely by the provider under the current contract. This has proven to be challenging for the MSSP, with a majority of participating providers unable to generate savings owing to the associated costs. In this paper, we propose a predictive analytics approach to redesigning the MSSP contract with the goal of better aligning incentives and improving financial outcomes from the MSSP. We formulate the MSSP as a principal–agent model and propose an alternate contract that includes a performance-based subsidy to partially reimburse the provider's investment. We prove the existence of a subsidy-based contract that dominates the current MSSP contract by producing a strictly higher expected payoff for both Medicare and the provider. We then propose an estimator based on inverse optimization for estimating the parameters of our model. We use a data set containing the financial performance of providers enrolled in the MSSP, which together accounts for 7 million beneficiaries and more than $70 billion in Medicare spending. We estimate that introducing performance-based subsidies to the MSSP can boost Medicare savings by up to 40% without compromising provider participation in the MSSP. We also find that the subsidy-based contract performs well in comparison with a fully flexible nonparametric contract.

Funding: This study was supported by the National Science Foundation [Grant CMMI-1450963] and by a Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarship.
Supplemental Material: Supplemental material is available at https://doi.org/10.1287/opre.2018.1821.

Keywords: contract design • structural estimation • healthcare policy • principal agent model • inverse optimization

1. Introduction

The United States spends more on healthcare than any other high-income nation in the world, both on a per capita basis and as a share of its gross domestic product (GDP). In 2016, total healthcare spending in the United States exceeded $3 trillion, equivalent to 17% of the U.S. GDP (OECD 2016). Approximately 20% ($600 billion) of U.S. healthcare spending is through Medicare, the United States' federally funded health insurance program for seniors and other qualifying individuals (CMS 2016a). An additional 17% of total healthcare spending is through Medicaid, which is targeted at low-income individuals and those with disabilities. Combined, government programs therefore account for 37% of all U.S. healthcare spending. Despite enormous spending on healthcare, the U.S. underperforms its international peers on many indicators of health quality (Starfield 2000, McGlynn et al. 2003), which suggests that high costs may be a consequence of *inefficiencies* in the healthcare system (Wennberg et al. 2002, Garber and Skinner 2008, Berwick and Hackbarth 2012). Moreover, although costs are already high, total healthcare spending—including Medicare—is projected to continue to climb at a rate of 5% per year over the next decade and outpace GDP growth by more than 1% (CMS 2016a).

The growing cost of healthcare presents a significant challenge for Medicare. The issue is further complicated by a misalignment of incentives between Medicare and providers: although Medicare bears the cost of healthcare delivery for its beneficiaries, providers may be disincentivized from delivering efficient, high-quality care under existing payment models (Rosenthal et al. 2004). For example, under traditional fee-for-service payment—whereby providers receive payments in proportion to

the volume and intensity of services provided—it may be profitable for a provider to overutilize diagnostic imaging (such as computed tomography scans) to generate additional Medicare payments, even if doing so does not necessarily improve patient outcomes (Hendee et al. 2010). Owing to the potential impact on total Medicare spending, correcting adverse incentives in healthcare delivery has become one of the central issues of healthcare reform (Milgate and Cheng 2006, Miller 2009, Wilensky 2013). As a consequence, Medicare payment models and provider incentive programs have recently received significant attention within the operations management community, with several studies focusing on analyzing and improving the efficacy of specific Medicare programs (Ata et al. 2012, Lee and Zenios 2012, Andritsos and Tang 2015, Gupta and Mehrotra 2015, Adida et al. 2016, Bastani et al. 2016, Guo et al. 2016, Zhang et al. 2016a).

The Medicare Shared Savings Program (MSSP) is a recent federal initiative administered by the Centers for Medicare and Medicaid Services (CMS) that encourages providers to cut costs by improving the efficiency of healthcare delivery. The MSSP has attracted significant attention in the health policy community owing to its potential impact on total spending, and as of 2017, it accounts for one-sixth of the Medicare population. However, the financial performance of providers enrolled in the MSSP indicates mixed results. In this paper, we propose an alternate contract for the MSSP and show that our proposed contract can lead to improved outcomes for both Medicare and participating providers. In the remainder of this section, we provide a brief overview of the MSSP and a summary of our contributions.

1.1. Overview: Medicare Shared Savings Program

The Medicare Shared Savings Program is a voluntary program that offers providers bonus payments for reducing the cost of providing care for Medicare beneficiaries, subject to satisfying certain quality standards. The goal of the MSSP is to correct the misalignment of incentives between Medicare and providers by making it financially viable for providers to improve the efficiency of healthcare delivery. To participate in the MSSP, a group of Medicare providers is required to form a cooperative entity referred to as an *Accountable Care Organization* (ACO). In contrast to the common practice of disaggregated providers treating patients independently, ACOs represent a shift toward an integrated care model, wherein a group of Medicare providers coordinates care for a well-defined beneficiary population, and group members are held jointly responsible for the quality of care delivered (Berwick 2011, Crosson 2011). The main focus of the ACO model of healthcare delivery is thus to improve the overall efficiency of providing healthcare through better coordination (e.g., by avoiding duplication of health services), with the aim of reducing healthcare spending while maintaining a high standard of care. As an example, Akira Health, Inc., is an ACO based in northern California that consists of 31 independent primary care providers (e.g., individual physicians or clinics), which together deliver care to 8,900 Medicare beneficiaries and account for $135 million in Medicare spending (Akira Health, Inc. 2017).

The defining feature of the MSSP is the establishment of *financial benchmarks* for each ACO, which are calculated on the basis of the ACO's historical spending and are risk adjusted according to attributes of the ACO's beneficiary population (Federal Register 2011). An ACO whose annual Medicare spending is less than its financial benchmark is eligible for a *shared savings* payment, which is a bonus payment made to the ACO in addition to the usual Medicare fee-for-service payment. The shared savings payment is proportional to the savings generated by the ACO (i.e., the difference between the benchmark and actual spending) so as to encourage the ACO to maximize savings. As of 2018, the Shared Savings Program has enrolled 561 ACOs, which together serve a total of 10.5 million Medicare beneficiaries, equivalent to one-sixth of the total Medicare population (CMS 2016a).[1]

Despite significant interest from providers, the MSSP faces two notable challenges. First, financial data released by CMS suggest that ACOs have struggled to cut costs: of the 392 ACOs participating in the MSSP as of 2015, only one-third generated enough savings to qualify for a shared savings payment (CMS 2017a). As a result, the reduction in total Medicare spending has been modest. In 2015, the total savings across all ACOs was approximately $430 million, which represents a 0.6% decrease in Medicare spending (CMS 2017a). Second, the low success rate among ACOs has made continued participation in the MSSP unattractive for many providers. A recent survey found that two-thirds of ACOs are uncertain whether they will continue participating in the Shared Savings Program, with less than 10% certain that they will remain enrolled (National Association of ACOs 2014). Because the success of the MSSP depends on voluntary participation by ACOs and a reduction in Medicare spending (Rosenthal et al. 2011), these challenges raise serious questions regarding the sustainability of the MSSP in its current form. Moreover, a major barrier that ACOs face with respect to cutting healthcare costs is the significant *investment* that must be made to improve the efficiency of healthcare delivery (Haywood and Kosel 2011). For example, an ACO may need to invest in new information technology to better coordinate care for its patients (Moore et al. 2011) or invest to increase the quality of care delivered so that costly readmissions are minimized (Anderson and Steinberg 1984). ACOs thus face a delicate balancing act: to cut costs and receive shared savings payments from Medicare, they must also *increase* spending to achieve the necessary efficiency gains.

1.2. Outline and Contributions

In this paper, we take a predictive analytics approach to redesigning the MSSP contract, with the goal of addressing the challenges currently faced by the MSSP. Specifically, we propose a simple and intuitive modification of the existing MSSP contract: a performance-based subsidy for an ACO's investment. The intention of the subsidy is twofold: first, to encourage ACOs to generate additional savings by offsetting the cost of efficiency improvements and, second, to boost total ACO payments so as to make the MSSP more attractive to current and prospective ACOs. Our analytical and empirical results suggest that the proposed subsidy scheme achieves both of these desirable outcomes.

We note here that a total overhaul of the MSSP would likely lead to the greatest improvement in Medicare savings. However, our perspective in this paper is that owing to the size of the program ($70 billion in Medicare spending) and the associated institutional inertia behind it, a straightforward adjustment to the existing contract is more realistic and stands a greater chance of being implemented in practice rather than a full redesign. For this reason, we retain much of the structure of the existing MSSP contract in our analysis and focus on the impact of incorporating an investment subsidy into the existing contract. For completeness, we also assess the potential impact of a full redesign of the MSSP contract and compare its performance with the subsidy-based contract.

Our analysis unfolds in two parts. First, we place the MSSP within a principal–agent framework and characterize the optimal contract. In the model, the ACO (agent) has the ability to make an investment to reduce spending on healthcare delivery and, in turn, earns bonus payments that depend on both the savings generated and parameters of the MSSP contract. The space of feasible contracts that Medicare may select from is defined by two parameters: a *shared savings rate*, which is the fraction of savings that the ACO receives as bonus payment, and a *subsidy rate*, which is the fraction of the ACO's investment that is reimbursed by Medicare. The solution to Medicare's optimal contracting problem is not obvious. A generous contract may provide a strong incentive for the ACO to reduce spending but will also increase Medicare's total payments to the ACO. Conversely, if the bonus payment offered to the ACO is too low, then the ACO may be insufficiently incentivized to generate any appreciable savings.

In the second part of our analysis, we design a new contract guided by financial performance data from 392 ACOs currently enrolled in the MSSP. We propose an estimator to infer parameters of the principal–agent model from the ACO data. In contrast to previous work on data-driven incentive design, which uses approaches such as linear or logistic regression (e.g., Lee and Zenios 2012, Yamin and Gavious 2013), our estimation procedure is in the spirit of *inverse optimization*, which refers to the inference of optimization model parameters from noisy solution data (Bertsimas et al. 2015, Aswani et al. 2018). Our approach is also distinguished from previous papers in that the key model primitive that we aim to estimate is itself a probability distribution. Using the estimated principal–agent model, we then solve for the optimal MSSP contract, which we formulate as a pure-integer optimization problem. Finally, we estimate the potential improvement in savings due to the subsidy by simulating ACO performance under both the existing and proposed contracts. Our main results are summarized as follows.

1.2.1. Characterization of the Optimal Contract. We first prove that under reasonable conditions, introducing a performance-based subsidy to the MSSP can increase total Medicare savings, despite the additional payments that must be made to the ACO. Because the ACO always benefits from subsidy payments, this result implies the existence of a subsidy-based contract that *dominates* all possible contracts within the current program. In other words, we show that introducing a subsidy can boost Medicare savings in addition to increasing ACO payments, making the MSSP more attractive to ACOs.

1.2.2. Estimate of Medicare Savings. We estimate that under the proposed contract, Medicare savings and ACO payments may increase by 43% and 17%, respectively, which supports the dominance result discussed above and provides evidence in favor of incorporating ACO investment subsidies into a revised MSSP contract. We also consider a complete redesign of the MSSP contract that replaces the shared-savings and subsidy components with a fully flexible nonparametric contract. As one would expect, we find that the nonparametric contract improves on the subsidy-based contract with respect to Medicare's savings. However, we find that the subsidy-based contract performs relatively well compared with the nonparametric contract, which suggests that a large share of the savings improvement associated with a full redesign of the MSSP might be attainable by subsidizing ACO investments.

1.2.3. Impact of Financial Benchmarks. Our estimation also reveals that ACOs with high benchmark expenditures are more likely to be effective at generating savings than ACOs with low benchmarks. On average, we find that ACOs with benchmark expenditures of less than $10,000 per beneficiary were unable to generate savings on a per-beneficiary basis, whereas ACOs with benchmarks greater than $14,000 reduced spending by $260 per beneficiary. This disparity in ACO performance may be explained by the fact that the financial benchmarks are calculated on the basis of the ACO's historical spending. As a consequence, an ACO with historically high spending may have relatively more "room for

improvement" than an ACO that has historically been cost efficient. Because ACO bonus payments increase in their savings, our finding suggest that Medicare should anticipate that ACOs with low benchmarks may drop out of the MSSP.

The remainder of this paper is organized as follows. In Section 2, we review related literature. In Section 3, we formulate the principal–agent model for the MSSP, provide an overview of the existing contract, and propose our subsidy-based contract. In Section 4, we formulate Medicare's optimal contracting problem and present our key analytical results. In Section 5, we consider variants of the optimal contracting problem, including the nonparametric contract. In Section 6, we present an inverse-optimization-based estimator for the principal–agent model and outline our empirical method. In Section 7, we present results from the structural estimation and discuss their policy implications with respect to the MSSP. We conclude in Section 8.

2. Related Literature

Our paper builds on a recent and growing body of work on healthcare contracts in the operations management literature. We also contribute to the health policy literature on accountable care organizations and the broader literature on incentive design.

2.1. Healthcare Contracts

Incentive problems in healthcare delivery have received significant attention in the operations management literature recently, in part owing to the focus on healthcare reform in the United States. A large share of this work has focused on principal–agent settings where the principal (e.g., Medicare) is required to design a payment model for an agent (e.g., provider) that yields socially beneficial outcomes, such as improvements to patient health or a reduction in healthcare spending. One of the first papers in the operations management literature to consider a healthcare contracting problem is by So and Tang (2000), who consider a setting where the principal reimburses an agent for drugs prescribed to patients. The authors focus on analyzing the agent's response to changes in a reimbursement policy that is tied to patient outcomes. Fuloria and Zenios (2001) consider a general problem in which an agent determines the intensity of treatment for a patient and the principal reimburses the agent for the services provided. Jiang et al. (2012) consider an optimal contracting problem in a general healthcare setting where the agent's decision is to allocate outpatient service capacity to different groups of patients, and the principal wishes to minimize service cost subject to constraints on agent performance. A common conclusion in all three of these papers is that linking provider reimbursement to patient health can lead to improved health outcomes. The contract we consider in this paper is different in that it depends primarily on the financial performance of the provider instead of the quality of healthcare delivered.

Several previous works have analyzed specific Medicare programs. Ata et al. (2013) consider Medicare's payment policies for hospice care. They highlight adverse incentives in the existing policies and propose an alternative payment model that corrects the misalignment of incentives. Gupta and Mehrotra (2015) formulate a game theoretic model for the Bundled Payments for Care Improvement Initiative, which is a new payment model in which Medicare providers receive a single payment for a collection of services provided to a beneficiary. Adida et al. (2016) and Guo et al. (2016) also analyze bundled payment models and compare their performance with traditional fee-for-service payment. Zhang et al. (2016a) propose a game theoretic model to study the behavior of hospitals under the recently created Hospital Readmissions Reduction Program, which is a mandatory program that penalizes hospitals that do not reduce readmissions below target levels. Their main result is a set of conditions under which a hospital would rather pay penalties than reduce readmissions. Andritsos and Tang (2015) compare the effect of different Medicare payment models (e.g., fee-for-service versus performance-based payment) on hospital readmissions in a setting where service is coproduced by both the provider and the patient. Bastani et al. (2016) propose a general principal–agent framework for pay-for-performance contracts and examine three Medicare programs as special instances. Jiang et al. (2016) and Savva et al. (2016) also consider the realignment of provider incentives in a more general setting and focus on the role of competition in reducing costs. To the best of our knowledge, the only existing work to consider the MSSP is by Zhang et al. (2016b), who analyze the impact of the MSSP on the use of computed tomography. Our paper is different in that we consider the impact of investment subsidies on the financial performance of ACOs and use observational data from the MSSP to estimate our model.

We are aware of two other papers that take a data-driven approach to designing incentives in a healthcare setting. The first is by Lee and Zenios (2012), who consider the problem of designing a payment model for Medicare's End Stage Renal Disease Program. A key methodological distinction with our work is that the formulation of the agent problem in Lee and Zenios (2012) lends itself naturally to the use of linear regression for estimating the model parameters. By contrast, the agent problem in our paper is a more general optimization problem and has no closed-form solution, which makes the use of linear regression nonviable. Instead, we take a maximum likelihood estimation approach, which results in requiring us to solve an inverse optimization problem (Aswani et al. 2018). A second distinction is that Lee and Zenios (2012) consider a linear contract, whereas

the current MSSP contract that we analyze is piecewise linear and discontinuous in the agent's output. The second related work is by Yamin and Gavious (2013), who consider incentives offered to patients to encourage them to receive influenza vaccinations. The authors formulate the vaccine problem within a game theoretic framework and use logistic regression to estimate the size of the incentive required to optimally vaccinate the population, using phone survey data. The context in the vaccine problem is markedly different from ours because the incentive is offered directly to patients rather than to healthcare providers, and the payment is provided as a lump sum rather than being a function of agent output or effort.

2.2. Health Policy

Although the MSSP is relatively recent (the first cohort of ACOs enrolled in 2012), it has received significant attention in the health policy literature owing its potential impact on healthcare spending. However, to date, quantitative analyses of ACO performance and the MSSP have been limited. McWilliams et al. (2016) analyze early ACO performance data and find that ACOs achieve minimal savings in their first year, suggesting a transition phase for ACOs once they enroll in the MSSP. They also find that ACOs consisting of independent primary care groups tend to save more than those integrated with hospitals. The authors carry out a similar analysis in McWilliams et al. (2013, 2015). Eddy and Shah (2012) develop a simulation model for ACO performance within the MSSP and find that the existing rules of the MSSP offer little incentive to ACOs to improve the quality of care delivered. Liu and Wu (2014) perform a simulation study that considers ACOs and patients as individual agents and focuses specifically on congential heart failure. There has also been significant qualitative discussion around the rules of ACO formation (Lieberman and Bertko 2011, Fisher et al. 2012) and the MSSP benchmarking methodology (Chernew et al. 2014, Douven et al. 2015). Our paper contributes to this literature by analyzing the MSSP from a modeling perspective and by being the first to examine the potential impact of ACO subsidies.

2.3. Incentive Design

Our paper builds on a rich and extensive literature on incentive design and principal–agent problems. For an overview of foundational work in the economics literature, we refer the reader to work by Hölmstrom (1979), Grossman and Hart (1983), and Hart and Holmström (1986). An overview of principal–agent problems is given by Gibbons (1998) and Laffont and Martimort (2009). Incentive design problems have also recently received significant attention in the operations management literature. Plambeck and Zenios (2000) propose a general framework for dynamic principal–agent problems based on a Markov decision process. Incentive problems have been considered in a wide variety of contexts in addition to healthcare, including software development (Whang 1992), finance (Grinblatt and Titman 1989, Raghu et al. 2003), sales (Chen 2000, DeHoratius and Raman 2007, Khanjari et al. 2013), project management (Chen et al. 2015), manufacturing (Balasubramanian and Bhardwaj 2004), and supply-chain management (Khouja and Zhou 2010, Guajardo et al. 2012, Lariviere 2015, Chen and Lee 2016). There is also a growing literature on the use of subsidies by a central planner in attaining socially desirable outcomes. Most of this work has focused on the use of subsidies to encourage product adoption, such as influenza vaccines (Chick et al. 2008, Arifoglu et al. 2012, Mamani et al. 2013), malaria drugs (Taylor and Xiao 2014, Levi et al. 2016), and renewable energy technologies (Ata et al. 2012; Cohen et al. 2015a, b; Chemama et al. 2019). We consider subsidies in a slightly different context, where they are used to offset the cost of agent effort rather than to reduce the purchase cost of a product.

3. Shared Savings Model

We begin by developing the principal–agent model for the MSSP, which we formulate as a single-period sequential game between the CMS ("Medicare") and a single ACO. The ACO provides healthcare to a beneficiary population at a cost that is entirely incurred by Medicare. The interaction proceeds in four steps. First, Medicare selects a contract that depends on the ACO's savings and investment. Second, the ACO observes the contract and invests in efficiency improvements to reduce spending. Third, the actual savings generated by the ACO is realized. Last, Medicare observes the actual savings and investment and pays (or receives a penalty from) the ACO according to the selected contract.

3.1. Preliminaries

The ACO is defined by two attributes: μ and θ. We refer to μ as the ACO's *benchmark*, which is known to Medicare, and θ as the ACOs *type*, which represents the ACO's private information and is unknown to Medicare. The benchmark serves as the primary reference point for determining whether the ACO has generated savings (if spending is less than μ) or losses (if spending is greater than μ). The benchmark is calculated on the basis of the ACO's historical spending in the years before joining the MSSP and is inflation adjusted to serve as an estimate of the cost of providing healthcare to the beneficiary population (Federal Register 2011). The type parameter θ governs the ACO's ability to generate savings. Let the interval $\Theta = [\underline{\theta}, \bar{\theta}]$ denote a continuum of possible ACO types. To reflect uncertainty in the ACO's type, let θ be a random variable supported on Θ, where $F(\cdot|\mu)$ and $f(\cdot|\mu)$ are the distribution and density, respectively, of an ACO with

benchmark μ. We assume that F and f are known to Medicare. The type distribution depends on the ACO's benchmark, meaning that the ACO's ability to generate savings may depend on its historical Medicare spending. We will often suppress dependence on μ in the notation. A type θ ACO can reduce spending by x by investing $c(x,\theta)$ into efficiency improvements, where $x \in [0,\bar{x}]$. The upper bound $\bar{x}$ is included for technical purposes and is without loss of generality. In general, if θ is large, we say the ACO is effective at generating savings and is a *high-type* ACO; conversely, if θ is small, then the ACO is ineffective at generating savings and is *low-type* ACO. This notion is formalized in Assumption 1.

Assumption 1. *Let the following assumptions hold:*
(i) $c(0,\theta) = 0$ *for all* θ,
(ii) $c(x,\theta)$ *is strictly convex and increasing in* x,
(iii) $c(x,\theta)$ *and* $(\partial/\partial x)c(x,\theta)$ *are decreasing in* θ, *and* $\lim_{\theta\to 0} c(x,\theta) = \infty$ *for any* x.

The assumption that $c(x,\theta)$ is strictly convex and increasing in x implies diminishing returns on investment because the marginal cost of generating an additional unit of savings increases with x. The assumption that $c(x,\theta)$ is decreasing in θ implies that high-type ACOs are more efficient at generating savings than low-type ACOs. Similarly, the assumption that $(\partial/\partial x)c(x,\theta)$ is decreasing in θ implies that the rate at which the return on investment diminishes is lower for high-type ACOs, which is consistent with the notion that high-type ACOs are more effective at generating savings. Examples of functional forms that satisfy Assumption 1 are x^2/θ and $(x/\theta)\log(x+1)$. Because of variations in the precise healthcare needs of the population, there may be uncertainty in the exact cost of delivering care. We account for this uncertainty through a random shock to the ACO's spending, denoted by ξ. Let $G(\cdot)$ and $g(\cdot)$ be the distribution and density of ξ, respectively. We assume that $G(\cdot)$ and $g(\cdot)$ are known to both Medicare and the ACO. Next, we impose the following conditions on the shock.

Assumption 2. *Let the following assumptions hold:*
(i) $\mathbb{E}[\xi] = 0$, and $\mathbb{E}[\xi^2] = \sigma^2 < \infty$.
(ii) *The shock density* g *is continuous, almost everywhere differentiable, unimodal, and symmetric around* 0.
(iii) *There exist constants* $\bar{g}, \bar{g}' < \infty$, *such that* $g(\xi) \le \bar{g}$, $g'_-(\xi) \le \bar{g}$, *and* $g'_+(\xi) \le \bar{g}$ *for all* ξ, *where* $g'_-(\xi)$ *and* $g'_+(\xi)$ *are left and right derivatives of* g, *respectively.*

Assumption 2 is fairly mild and admits a large class of probability distributions, such as the Gaussian, Laplace, and logistic distributions. Statement (iii) simply means that the density and its derivative are bounded and is stated with respect to the left and right derivatives of g to permit distributions that are not differentiable everywhere (e.g., the Laplace density). The shock distribution is independent of the ACO's benchmark. This assumption allows us to avoid overfitting by reducing the number of parameters in our estimation procedure (discussed in Section 6). We obtain a strong model fit despite this simplifying assumption.

We can now define the ACO's actual spending on delivering healthcare, or the *delivery cost*, as $D = \mu - x - \xi$. It may be convenient to interpret D as demand for healthcare that must be satisfied by the ACO. Because, in our setting, demand represents the healthcare needs of the patient population, we model it as being exogenous to costs. Because $\mathbb{E}[\xi] = 0$, the expected delivery cost if the ACO makes no investment (implying $x = 0$ by Assumption 1) is given by $\mathbb{E}[D] = \mu$. We can now define the *savings* generated by the ACO as the difference between the benchmark and actual delivery costs as $y = \mu - D$. Because $D = \mu - x - \xi$, we can write the savings equivalently as

$$y = x + \xi.$$

Therefore, the actual savings generated by an ACO is equal to its reduction in spending plus an exogenous shock. Let $\omega(\cdot|x)$ denote the density function for the savings given the ACO's action x. Note that the actual savings y is random owing to the shock and that $\mathbb{E}[y] = x$. We also assume throughout that $\mathbb{P}[\mu - \bar{x} - \xi < 0] = 0$, meaning that the demand is always nonnegative. This assumption can be ensured by truncating the distribution of the shock variable ξ (and holds trivially in practice, given that healthcare costs cannot be driven to 0). Next, we define the contract that determines the ACO's shared savings payment, which depends on the realized savings, y.

3.2. Baseline Contract: Shared Savings

In the MSSP, an ACO enters a "one-sided" contract for an initial agreement period (usually three years), after which it is transitioned to a "two-sided" contract. The one-sided contract is risk free for the ACO: it receives payments from Medicare for generating savings but is not penalized if spending exceeds the benchmark. The one-sided contract depends on three parameters: the *shared savings rate* $\alpha \in \mathcal{A}$, a minimum savings threshold h, and a fixed savings cap C_u. The shared savings payment received by the ACO in the one-sided contract is then given by

$$r_+(y,\alpha) = \begin{cases} 0, & y \in (-\infty, h], \\ \alpha y, & y \in (h, C_u], \\ \alpha C_u, & y \in (C_u, \infty). \end{cases} \tag{1}$$

The shared savings rate α represents the fraction of savings that the ACO receives as payment if the realized savings y is positive. The threshold h is the minimum savings required for the ACO to receive a payment from Medicare. The minimum savings threshold accounts for natural variation in healthcare costs by ensuring that any

observed savings are "real" (i.e., due to ACO effort and not chance; Federal Register 2011). The savings cap C_u reduces the risk to Medicare due to the shock and the information assymmetry by protecting Medicare from making an excessively large shared savings payment to the ACO.

In the two-sided contract, both savings and losses are shared with Medicare. In addition to receiving a shared savings payment for generating positive savings, the ACO pays a penalty of $(1-\alpha)y$ to Medicare if $y < -h$, that is, if the excess spending above the benchmark is greater than the threshold h. Similar to the one-sided case, the maximum penalty that can be paid by the ACO is capped at $(1-\alpha)C_\ell$. Writing the penalty terms as negative payments, the payment received by the ACO in the two-sided contract is then given by

$$r(y,\alpha) = \begin{cases} (1-\alpha)C_\ell, & y \in (-\infty, C_\ell], \\ (1-\alpha)y, & y \in (C_\ell, -h], \\ 0, & y \in (-h, h], \\ \alpha y, & y \in (h, C_u], \\ \alpha C_u, & y \in (C_u, \infty). \end{cases} \tag{2}$$

Figure 1 shows a schematic of the one- and two-sided contracts of the MSSP. We largely focus on the two-sided contract in this paper because it is the intended long-term contract of the MSSP. The one-sided contract is most relevant in Section 6 because the financial data that we use for estimation correspond to ACO performance under the one-sided contract.

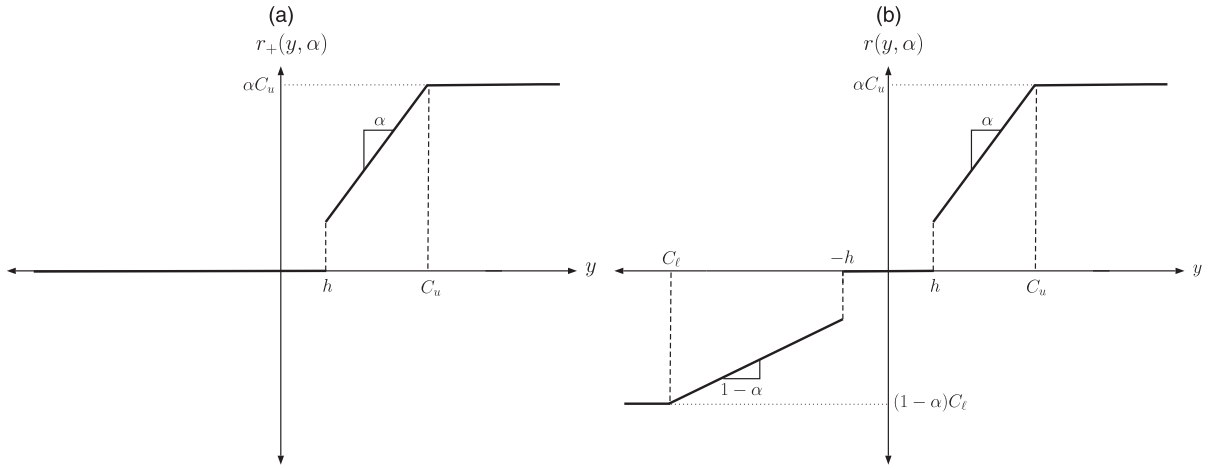
3.3. Proposed Contract: Shared Savings + Performance-Based Subsidy

The MSSP in its current form is entirely performance based, in that it rewards the ACO for generating savings but is otherwise agnostic to the ACO's investment. The presence of the random shock under this payment model is risky for the ACO because it may be penalized for exceeding the benchmark despite investing in efficiency improvements. Here we define a new contract that partially mitigates the ACO's risk by providing a subsidy for the investment in addition to the the existing shared savings payment. In this sense, our proposed contract is both effort based and performance based. Further, as discussed in Section 1, our inclusion of a subsidy component in the MSSP contract is related to evidence that suggests that the investments required by ACOs to improve efficiency can be a barrier to the generation of savings (CMS 2016c). Moreover, our approach of contracting on both the ACO's action and its performance is inspired by a well-established body of literature that suggests that contracting on both action and outcome can improve outcomes (Harris and Raviv 1979; Hölmstrom 1979; Shavell 1979a, b). As we discuss later in this section, the key consequence of the proposed contract is that under reasonable conditions, it can dominate the existing contract by generating a higher payoff for *both* Medicare and the ACO. Under the proposed contract, Medicare observes the ACO's investment and provides a subsidy proportional to the investment, in addition to the shared savings payment. Let $\beta \in \mathcal{B}$ be the *subsidy rate* selected by Medicare. The subsidy payment is then given by

$$s(y,\beta) = \begin{cases} 0, & y \in (-\infty, h], \\ \beta c(x,\theta), & y \in (h, \infty). \end{cases} \tag{3}$$

Similar to the shared savings payment, the subsidy payment is also performance based, in the sense that the ACO only receives it if the savings is positive and exceeds the minimum savings threshold h. Note that the subsidy payment $\beta c(x,\theta)$ depends on the private information θ, which is a priori unknown to Medicare. In line with standard mechanism design theory, in

Figure 1. Shared Savings Payment Functions Under One-Sided and Two-Sided Contracts



One-sided contract. Two-sided contract.

Section 4.2 we formulate Medicare's optimal contracting problem in a manner that incentivizes ACOs to truthfully report their type θ, which enables Medicare to contract directly on the ACO's investment.

We also highlight here that the subsidy scheme presented in this paper is different from the ACO Investment Model (AIM) that was recently created by the CMS (CMS 2016c). The most significant difference is that the subsidy payments offered to ACOs through the AIM are later deducted from shared savings payments, which effectively makes the AIM a loan program. We instead consider an entirely separate incentive that exists in addition to the shared savings payment. Further, the AIM is targeted at a subset of ACOs that meet a well-defined criterion (e.g., those that are in rural areas and excludes hospitals), whereas we consider a more general program that can apply to any ACO.

3.4. Discussion of Modeling Assumptions

Before analyzing Medicare's contracting problem, we first discuss some of the key assumptions and limitations of our model.

3.4.1. Quality. One might reasonably expect that an ACO could decrease the quality of healthcare delivered (e.g., by cutting services) as a way to reduce spending and generate bonus payments within the MSSP. However, we assume throughout our analysis that quality of care is fixed and that the ACO only decides on the size of the investment. In other words, our model does not permit the ACO to decrease quality of care to generate savings. This assumption is supported by ACO performance data and the existing regulations of the MSSP. Data released by the CMS show that quality has actually *improved* under the MSSP—for example, the average ACO quality score increased by approximately 15% from 2014 to 2015 (CMS 2016b). Moreover, the regulations of the MSSP require Medicare to verify that any savings generated by an ACO are not due to quality reductions. Medicare upholds quality through close monitoring of ACOs (e.g., to ensure that they do not avoid at-risk patients or underuse healthcare services; CMS 2016d), and failure to meet quality standards or comply with Medicare monitoring may jeopardize an ACO's participation in the MSSP (Federal Register 2011). As a consequence, ACOs have neither the incentive nor the ability to use quality as a lever to generate shared savings payments. We therefore focus our analysis on the ACO's investment behavior.

3.4.2. Benchmark Independence and ACO Size. In our model, the ACO's optimal savings $x(\theta)$ depends explicitly on α, β, and θ but does not depend on the benchmark μ when θ is held fixed. To capture the effect of the ACO benchmark on the savings generated, we allow the distribution over the parameter θ to depend on the ACO's benchmark in our empirical analysis in Section 6. We also do not explicitly incorporate ACO size (i.e., number of assigned beneficiaries) into the model and instead perform our analysis at the beneficiary level. This normalization does not change the analytical results presented in this section but is useful for simplifying the estimation in Section 6. A cursory look at the data suggests a weak relationship between ACO size and savings (a correlation coefficient of $\rho = 0.08$), which supports this normalization. Although we normalize for beneficiaries for estimation purposes, we explicitly incorporate ACO size when estimating ACO performance under the proposed contract in Section 6.

3.4.3. Single ACO. Although the MSSP includes many participating ACOs, for our analytical results, we consider the interaction between Medicare and a single ACO. This is because an ACO's payment depends only its own savings, benchmark, and investment and does not depend on the performance of other ACOs. Moreover, ACOs are prohibited by legislation from colluding or sharing information with each other (Federal Register 2011). As a consequence, the key insights can be obtained by focusing on the contracting problem between Medicare and one ACO.

3.4.4. Single Period. In practice, the financial benchmarks for ACOs are updated annually, on the basis of the ACO's most recent three years of spending and the growth rate in national healthcare spending. The MSSP might therefore be viewed as a multiperiod problem with dynamically updating spending benchmarks. However, because our aim in this paper is to analyze the potential impact of investment subsidies, it suffices to restrict our attention to a single-period model. We note here that additional insights may be gained by considering a multiperiod model for the MSSP. We take a first step toward this extension in Section EC.4 of the online supplemental material.

4. Model Analysis

In this section, we consider the ACO's investment behavior and Medicare's optimal contracting problem under the existing and proposed MSSP contracts. In Section 4.1, we present the ACO's investment problem. In Section 4.2, we present Medicare's contracting problem and analytical results regarding the optimal contract structure. In Section 4.3, we consider a variant of the optimal contracting problem whereby ACO participation is determined endogenously.

4.1. ACO Investment

We begin our analysis by considering the investment problem faced by the ACO. Medicare providers typically generate profit from delivering healthcare to beneficiaries. Let the profit associated with spending D be γD, where $\gamma > 0$ is the profit margin. The expected service-related

profit is then given by $\mathbb{E}[\gamma D] = \gamma(\mu - x)$. Because μ is a constant, we focus on the ACO's profit loss due to a reduction in spending, which is simply γx.[2] The ACO's *expected payoff* associated with a savings level of x can now be written as

$$u(x,\alpha,\beta,\theta) = \int_{-\infty}^{\infty} (r(y,\alpha) + s(y,\beta))\omega(y|x)dy - \gamma x - c(x,\theta). \tag{4}$$

The payoff function $u(x,\alpha,\beta,\theta)$ is the sum of the expected shared savings and subsidy payments from Medicare minus the profit loss and investment. We assume that the ACO wishes to maximize expected payoff. The ACO's *optimal savings* is then given by

$$x(\theta) = \operatorname*{argmax}_{x\in[0,\bar{x}]} \; u(x,\alpha,\beta,\theta), \tag{5}$$

where $\bar{x}$ is the maximum savings the ACO can achieve (e.g., $\bar{x}$ can be trivially set to μ). In general, x depends on α, β, and σ as well, although we suppress dependence on these parameters for conciseness when it is clear from context. We assume that the ACO participates in the MSSP if $U(\alpha,\beta,\theta) \geq 0$.[3] The model naturally generalizes to account for a fixed cost associated with enrolling in the MSSP, which can be represented by subtracting a constant from $u(x,\alpha,\beta,\theta)$. Including this fixed cost does not fundamentally alter our results, so we assume it to be 0. We now turn our attention to analyzing the ACO's behavior, starting with the following assumption.

Assumption 3. *The inequality*

$$\bar{\alpha}(h\bar{g}' + \bar{g}) + 2\bar{\beta}\bar{g}(\partial/\partial x)c(x,\theta) \leq (1-\bar{\beta})(\partial^2/\partial x^2)c(x,\theta)$$

holds for all $x \in [0,\bar{x}]$ *and* $\theta \in \Theta$.

We now make the following observation.

Lemma 1. *Under Assumption 3, the ACO payoff function* $u(x,\alpha,\beta,\theta)$ *is strictly concave over* $[0,\bar{x}]$.

Assumption 3 is a technical condition that faciliates our analysis by guaranteeing that the ACO's payoff function has a unique maximizer and can be shown to hold under parameter values from our data set (see Section EC.1 in the online supplemental material). We emphasize that strict concavity in the ACO's payoff function is not crucial, and our results remain valid for any payoff function with a unique maximizer (i.e., nonconvex unimodal functions). We now define a quantity that will be useful in the remainder of this section.

Definition 1. Define $\theta_{\alpha,\beta} = \inf\{\theta \in \Theta | x(\theta) > 0 \, under \, \alpha,\beta\}$.

The threshold $\theta_{\alpha,\beta}$ represents the lowest type ACO that would generate a strictly positive savings given the shared savings rate α and subsidy rate β (note that $\theta_{\alpha,0}$ may not exist if α is very small; in general, we assume that the minimum shared savings rate in $\mathcal{A}$ is large enough that $\theta_{\alpha,0}$ exists for all $\alpha \in \mathcal{A}$). We now introduce two lemmas that characterize the ACO's behavior and will be useful for our main results in the next section. The first of these relates the savings generated by an ACO to its type.

Lemma 2. *For any* $\alpha \in \mathcal{A}$ *and* $\beta \in \mathcal{B}$, *the ACO's optimal savings* $x(\theta)$ *is nondecreasing over* Θ *and strictly increasing over* $[\theta_{\alpha,\beta},\bar{\theta}]$.

Lemma 2 states that an ACO with high investment efficacy will generate higher savings. This is not a surprising result—because it is less costly for a high-type ACO to achieve a fixed savings level x, one might expect that it would be optimal for a high-type ACO to save more than a low-type ACO. The parameter $\theta_{\alpha,\beta}$ is the threshold beyond which ACO savings are guaranteed to be strictly increasing in type. This threshold exists because for low-type ACOs where $x = 0$, a small increase in investment efficacy may not be sufficient to make it profitable for the ACO to generate positive savings, and thus x may remain at 0. However, if $x > 0$, then even slight improvements in the ACO's investment efficacy will lead to higher savings. The next lemma shows that offering a subsidy can boost ACO payoff while also incentivizing the ACO to generate higher savings.

Lemma 3. *For any* $\alpha \in \mathcal{A}$ *and* $\theta \in \Theta$, *the ACO's optimal savings* $x(\theta)$ *and payoff* $U(\alpha,\beta,\theta)$ *are nondecreasing in* β. *For any* α *and* $\theta \in [\theta_{\alpha,0},\bar{\theta}]$, *the savings* $x(\theta)$ *and* $U(\alpha,\beta,\theta)$ *are strictly increasing in* β.

Intuitively, the subsidy effectively increases the ACO's investment efficacy because it makes it less costly for the ACO to generate a fixed saving of x. As a consequence, the ACO's investment and optimal savings level increases with β. Further, the ACO's optimal payoff $U(\alpha,\beta,\theta)$ also increases with β. This suggests that introducing a subsidy may help improve participation by boosting total ACO payments.

4.2. Medicare's Optimal Contract

For a given ACO type θ, Medicare's total expected spending is given by the sum of the delivery cost and the shared savings and subsidy payments made to the ACO. Because the expected delivery cost is $\mathbb{E}[D] = \mu - x$, and μ is a constant, we can formulate Medicare's problem equivalently as one of maximizing savings. We thus write *Medicare's savings* as the ACO's savings minus the payments made to the ACO:

$$v(x,\alpha,\beta,\theta) = x - \int_{-\infty}^{\infty} (r(y,\alpha) + s(y,\beta))\omega(y|x)dy. \tag{6}$$

We now formulate the contracting problem faced by Medicare. We assume that the the ACO's type θ is only known to be drawn from the known distribution $f(\theta)$. In this setting, Medicare's contracting problem is to maximize savings by designing a menu of contracts

$(\alpha(\theta), \beta(\theta))$, $\theta \in \Theta$, where $\alpha(\theta) \in \mathcal{A}$ and $\beta(\theta) \in \mathcal{B}$ for all $\theta \in \Theta$. An ACO that reports its type to be θ then receives the contract $(\alpha(\theta), \beta(\theta))$. We assume that Medicare cannot adjust the parameters of the contract after the ACO reports its type. For convenience, we also define $U(\alpha, \beta, \theta) = u(x(\theta), \alpha, \beta, \theta)$ and $V(\alpha, \beta, \theta) = v(x(\theta), \alpha, \beta, \theta)$ to be the ACO's profit and Medicare's savings under the ACO's optimal savings $x(\theta)$. Medicare's optimal contracting problem is given by

$$\underset{\alpha(\theta), \beta(\theta)}{\text{maximize}} \quad \int_{\theta \in \Theta} V(\alpha(\theta), \beta(\theta), \theta)\, f(\theta) d\theta \tag{7a}$$

$$\text{subject to} \quad U(\alpha(\theta), \beta(\theta), \theta) \geq \bar{u}(\theta), \quad \theta \in \Theta, \tag{7b}$$

(OC-I)

$$U(\alpha(\theta), \beta(\theta), \theta) \geq U(\alpha(\theta), \beta(\theta), \theta') \quad \theta, \theta' \in \Theta, \tag{7c}$$

$$\alpha(\theta) \in \mathcal{A}, \quad \theta \in \Theta, \tag{7d}$$

$$\beta(\theta) \in \mathcal{B}, \quad \theta \in \Theta. \tag{7e}$$

Let $(\alpha^*(\theta), \beta^*(\theta))$, $\theta \in \Theta$ denote optimal contract parameters attained as a solution to OC-I. The objective function (7a) represents Medicare's expected savings, where the expectation is taken over both the shock and type. The constraint (7b) is a participation constraint that ensures that the ACO receives a payoff of at least $\bar{u}(\theta)$. Participation constraints are standard in the mechanism design literature and are also referred to as *individual rationality* constraints (e.g., see Laffont and Martimort 2009 and Borgers et al. 2015). This constraint is particularly important in the context of the MSSP because voluntary participation by a large number of ACOs is crucial for the success of the program (Rosenthal et al. 2011) (for completeness, we later consider a contracting scheme in which this participation constraint is relaxed). Constraints (7c) are *incentive compatibility* constraints that enable Medicare to accurately elicit the ACO's private information θ (Laffont and Martimort 2009). The interpretation of constraints (7c) is that the payoff that the ACO enjoys from truthfully reporting its type θ to Medicare must be at least as high as the payoff it receives from declaring any other type θ'. These constraints ensure that the feasible set of contracts is restricted to those that allow the ACO's type θ—and by extension, its optimal investment $c(x(\theta), \theta)$—to be elicited by Medicare.[4] Note also that OC-I is a bilevel program, because the constraint sets depend on the optimal solution of the ACO problem $x(\theta)$ [see Dempe (2002) for an overview of bilevel programming]. Bilevel programs are typically nonconvex, so the uniqueness of $\alpha^*(\theta)$ and $\beta^*(\theta)$ is generally not guaranteed.

Next, we present three analytical results that characterize the optimal contract. First, let the following assumption hold in the remainder of this section.

Assumption 4. *The inequality* $(\partial/\partial x)R(x, \alpha) < 1$ *holds for all* $x \in [0, \bar{x}]$.

This assumption is relatively mild and required for technical purposes. Intuitively, we would expect Assumption 4 to hold at larger values of x because the expected shared savings payment is sublinear at higher savings levels (owing to $\alpha < 1$). However, there can exist values of h and σ such that Assumption 4 is violated for small x, owing to the discontinuity of the shared savings payment at $y = h$. This assumption is validated by our data set (see Section EC.1 of the online supplemental material). Our first result regarding the optimal contract focuses on low-type ACOs.

Proposition 1. *There exist* $\psi > 0$ *and* $\theta_0 > 0$ *such that if* $C_u \geq \psi$ *and* $|C_\ell| \geq \psi$, *then* $\alpha^*(\theta) \geq 1/2$ *and* $\beta^*(\theta) = 0$ *for all* $\theta \leq \theta_0$.

The condition that C_u and $|C_\ell|$ be sufficiently large simplifies the analysis but is not restrictive.[5] To further illustrate Proposition 1, recall that the optimal savings $x(\theta)$ of a very low-type ACO will be small (Lemma 2). As a consequence, the shared savings rate α must be sufficiently large to guarantee the ACO a nonnegative payoff. From the ACO performance data, we observed that the average quality-adjusted shared savings rate that ACOs received in 2015 was 0.55, with approximately 15% of ACOs receiving a shared savings rate of less than 0.5. In light of Proposition 1, this suggests that a large share of ACOs may be at risk of dropping out of the MSSP once they transition to the two-sided contract. The next result characterizes the dependence of optimal contract parameters $\alpha^*(\theta)$ and $\beta^*(\theta)$ on θ.

Proposition 2. *Suppose that* $\Theta = \{\theta_L, \theta_H\}$, *where* $\theta_H > \theta_L > 0$. *There exists* $\delta > 0$ *such that if* $\theta_H - \theta_L \geq \delta$, *then* $\alpha^*(\theta_L) \geq \alpha^*(\theta_H)$ *and* $\beta^*(\theta_L) \leq \beta^*(\theta_H)$.

Proposition 2 states that it is optimal for Medicare to offer a high-type ACO a higher subsidy rate and a lower shared savings rate compared with a low-type ACO. The condition that the ACO types be sufficiently far apart (as represented by the condition that $\theta_H - \theta_L \geq \delta$ for some $\delta > 0$) is included because the generality of $g(\xi)$ and $c(x, \theta)$, in addition to the absence of a closed-form expression for $x(\theta)$, precludes obtaining stronger monotonicity results regarding $\alpha^*(\theta)$ and $\beta^*(\theta)$. To see the intuition behind Proposition 2, note that a very high-type ACO can generate a large savings with a relatively small investment. Note also that the resulting subsidy payment associated with a subsidy rate of β is proportional to the ACO's investment. As a result, it is worthwhile for Medicare to provide a large subsidy rate to a high-type ACO because the actual subsidy payment will be small relative to the increase in savings generated by the ACO. The converse is true for low-type ACOS. Our final result in this section compares the outcomes of the optimal nonsubsidized[6] and subsidy-based contracts.

Proposition 3. *Let $U_0^*(\theta)$ and $U_s^*(\theta)$ be a type θ ACO's payoff under the optimal nonsubsidized and subsidy-based contracts, respectively. Let $V_0^*(\theta)$ and $V_s^*(\theta)$ be the associated Medicare savings. Then there exists $\theta_s > 0$ such that $U_s^*(\theta) > U_0^*(\theta)$ and $V_s^*(\theta) > V_0^*(\theta)$ for all $\theta \geq \theta_s$.*

Proposition 3 states that for sufficiently high-type ACOs, the subsidy-based contract dominates the existing shared savings only contract, in the sense that it produces a strictly higher payoff for both Medicare and the ACO. The condition that the ACO type is at least θ_s is necessary to ensure that the ACO is effective enough at generating savings for it to be worthwhile for Medicare to provide an additional payment in the form of a subsidy. Conversely, if the ACO type is too low, then it may be the case that the additional savings generated by the ACO is less than the subsidy payment itself, which would make it suboptimal for Medicare to offer a subsidy of any size. From our numerical results, we find that this type threshold θ_s is relatively low (see Section 7). To illustrate the need for Assumption 4 in Proposition 3, recall that increasing β from 0 to a positive value increases the ACO savings x (Lemma 3) and thus increases the expected shared savings payment as well. Now consider the case in which σ is very close to 0 so that $y \approx x$. Because $r(y, \alpha)$ contains a jump discontinuity at $y = h$, a small increase in x may lead to a relatively large increase in $r(y, \alpha)$ if x is close to h before the subsidy. Assumption 4 ensures that the increase in the expected shared savings payment is not greater than the increase in x itself. The key implication of Proposition 3 is that Medicare can generate additional savings using the subsidy *without* jeopardizing ACO participation through lower payments.

5. Endogenous ACO Participation and Nonparametric Contract

In this section, we present three alternate models for Medicare's optimal contracting problem. In the first model, the ACO's participation is determined endogenously instead of being enforced as a constraint (as in OC-I). This relaxation of the participation constraint may further improve Medicare's savings compared with formulation OC-I because it allows Medicare to shed inefficient ACOs from the program. This alternate formulation is given by

$$\underset{\alpha(\theta),\beta(\theta)}{\text{maximize}} \int_{\theta\in\Theta} \left(V(\alpha(\theta),\beta(\theta),\theta)\cdot \mathbf{1}_{\{U(\alpha(\theta),\beta(\theta),\theta)\geq 0\}}\right) f(\theta)d\theta \tag{8a}$$

$$\begin{aligned} &(\text{OC}-\text{II})\\ &\text{subject to} \quad U(\alpha(\theta),\beta(\theta),\theta) \geq U(\alpha(\theta),\beta(\theta),\theta') \\ &\qquad\qquad\qquad\qquad\qquad\qquad \theta,\theta' \in \Theta, && (8b)\\ &\qquad\qquad \alpha(\theta) \in \mathcal{A}, \quad \theta \in \Theta, && (8c)\\ &\qquad\qquad \beta(\theta) \in \mathcal{B}, \quad \theta \in \Theta. && (8d) \end{aligned}$$

Note that formulation OC-II does not include the participation constraint given in OC-I. Instead, we introduce the indicator $\mathbf{1}_{\{U(\alpha(\theta),\beta(\theta),\theta)\geq \bar{u}(\theta)\}}$ to the objective function, which equals 1 if a type θ ACO receives a payoff of at least $\bar{u}(\theta)$ under contract $(\alpha(\theta), \beta(\theta))$. Including this indicator ensures that the associated Medicare savings are only counted if the ACO participates under the given contract.

The second alternate model we consider is a nonparametric contract that does not involve separate shared savings or subsidy payments. Instead, the contract takes the form of a general payment schedule $\rho(x)$ that depends on the ACO's savings level x. Because $\rho(x)$ can be a general function, this nonparametric contract can be viewed as the most flexible possible contract for the MSSP. Within this framework, a type θ ACO's profit under payment $\rho(x)$ and savings level x is given by

$$u(x,\rho,\theta) = \rho(x) - \gamma x - c(x,\theta), \tag{9}$$

and Medicare's associated savings is given by $y - \rho$, which, in expectation, is simply

$$v(x,\rho) = x - \rho(x). \tag{10}$$

By the well-known *revelation principle* (Myerson 1981), it is sufficient for Medicare to restrict attention to designing a menu of contracts $(x(\theta), \rho(\theta))$, $\theta \in \Theta$, that maps a savings level and payment to each ACO type. Under incentive compatibility, an ACO that reports type θ finds it optimal to generate savings $x(\theta)$, for which it receives payment $\rho(x(\theta))$. With a slight abuse of notation, we again let $U(\cdot)$ and $V(\cdot)$ represent the ACO's profit and Medicare's savings under a type θ ACO's optimal savings, respectively. The optimal nonparametric contract is then given by the solution to

$$\begin{aligned} &\underset{x(\theta),\rho(\theta)}{\text{maximize}} && \int_{\theta\in\Theta} V(x(\theta),\rho(\theta),\theta)\, f(\theta)d\theta && (11a)\\ &\text{subject to} && U(x(\theta),\rho(\theta),\theta) \geq \bar{u}(\theta), \quad \theta \in \Theta, && (11b)\\ &(\text{OC}-\text{III}) && U(x(\theta),\rho(\theta),\theta) \geq U(x(\theta),\rho(\theta),\theta')\ \theta, \\ &&&\qquad\qquad \theta' \in \Theta, \\ &&&\qquad\qquad && (11c)\\ &&& x(\theta) \geq 0, \quad \theta \in \Theta, && (11d)\\ &&& \rho(\theta) \geq 0, \quad \theta \in \Theta. && (11e) \end{aligned}$$

Because the contracting problem OC-III represents the most flexible contract, it can be used as a theoretical benchmark against which to measure the impact of the subsidy-based contract. As with the subsidy-based contract, Medicare's optimal contracting problem OC-III can be formulated as an integer optimization problem (see Section EC.2 of online supplemental material). The third alternate model we present combines the salient features of the preceding two: it represents the optimal

nonparametric contract in a setting where ACO participation is also endogenous:

$$\underset{x(\theta),\rho(\theta)}{\text{maximize}} \int_{\theta\in\Theta} V(x(\theta),\rho(\theta),\theta)\,\mathbf{1}_{[U(x(\theta),\rho(\theta),\theta)\geq 0]} f(\theta)d\theta \tag{12a}$$

(OC – IV)

$$\text{subject to } U(x(\theta),\rho(\theta),\theta) \geq U(x(\theta),\rho(\theta),\theta') \quad \theta,\theta'\in\Theta, \tag{12b}$$

$$x(\theta)\geq 0, \quad \theta\in\Theta, \tag{12c}$$

$$\rho(\theta)\geq 0, \quad \theta\in\Theta. \tag{12d}$$

We have thus far formulated four contracting problems for Medicare, given by OC-I, OC-II, OC-III, and OC-IV. In the remainder of this paper, we outline an empirical approach to estimating the potential performance of each of these four contracts and assess their effectiveness with respect to the status quo MSSP contract.

6. Estimation

Typically, in the mechanism design literature, the distribution over agent types is common knowledge to the principal. In practice, however, this distribution is unlikely to be known a priori. Our aim in this section is to use ACO financial data made available by the CMS to estimate the type density $f(\theta)$, which is an important input to identifying the optimal MSSP contract.[7] We also estimate the variance of the shock σ^2 from the data. We then use these estimates to solve for the optimal parameters $\alpha(\theta)$ and $\beta(\theta)$ under the proposed contract. Last, we estimate the performance of ACOs in the MSSP under the existing and proposed contracts, which allows us to estimate the improvement in Medicare savings that might result from introducing a performance-based subsidy to the MSSP.

6.1. Data

We use a data set made publicly available by the CMS (CMS 2017a). The data set contains the number of Medicare beneficiaries, benchmark expenditures, actual expenditures, and quality score for 392 ACOs participating within the MSSP in 2015. A summary of the data set is given in Tables 1 and 2. The ACOs in our data set represent 7.27 million Medicare beneficiaries across the United States. In 2015, this group of ACOs was assigned an aggregate spending benchmark of \$73.30 billion and had an actual spending of \$72.87 billion, representing a \$430 million net reduction in total Medicare spending. Although the total savings was \$430 million, the total shared savings payment earned by ACOs was \$645 million, resulting in a net loss of \$215 million for Medicare. This loss is due to the fact that as of 2015, all participating ACOs were under an initial three-year agreement in which the one-sided contract was in effect, meaning that ACOs were rewarded for generating savings but not penalized for exceeding their assigned benchmarks. Note that the mean per-beneficiary savings by an ACO in the study cohort was \$101, which is approximately 1% of the mean benchmark. The standard deviation of the per-beneficiary ACO savings was 80, suggesting substantial variation in ACO performance. Figure 2 illustrates the distribution in the number of beneficiaries, spending benchmarks, and savings for all 392 ACOs.

Let μ_i, b_i, y_i, and θ_i represent the benchmark, number of beneficiaries, actual savings, and type of the ith ACO, respectively. In practice, μ_i, b_i, and y_i are directly observed by Medicare, but θ_i represents private information. To maximize Medicare's savings, we would ideally estimate $\theta_i,\ldots,\theta_n$ from the data and then identify the optimal contract to offer to each of the n ACOs, according to their types. However, this approach is not viable because limited data are available for each ACO (and in many cases, only a single year). As a consequence, the type parameters $\theta_i,\ldots,\theta_n$ cannot be estimated directly. Under appropriate assumptions, however, we may instead estimate the distribution over ACO types, $f(\theta)$, which enables us to solve for the optimal contract. For our numerical experiments, we select a random subset of 275 ACOs (70% of the data) to estimate the model and validate the resulting model fit against an out-of-sample data set containing the performance of the remaining 117 ACOs (30% of the data).

6.2. Model Parameterization

We have so far imposed minimal assumptions on the ACO type and spending shock distributions. Our goal now is to estimate these distributions from the ACO performance data in a manner that is both tractable and captures the dependence of an ACO's type on its benchmark. Here we outline a parameterization of our model to be used in the estimation. The model primitives to be specified are the ACO conditional type distribution $f(\theta|\mu)$, the shock distribution $G(\xi)$, and the ACO's investment function $c(x,\theta)$. We assume that ACO types are distributed on the type interval $\Theta=[1,1\times 10^4]$ according to a mixture of (truncated) exponential

Table 1. Performance of ACOs in Medicare Shared Savings Program in 2015 Under One-Sided Contract

Variable	Value
ACOs	392
Total Medicare beneficiaries	7.27 million
Total benchmark expenditures	\$73.30 billion
Total actual expenditures	\$72.87 billion
Total savings	\$430 million
ACO shared savings payments	\$645 million
Medicare savings	–\$215 million

Table 2. ACO Summary Statistics.

	Mean (standard deviation)	Median	Interquartile range	Range
Beneficiaries (per ACO)	18,547 (18,508)	12,545	7,954–21,286	513–149,633
Benchmark ($ per beneficiary)	10,403 (2,360)	9,863	8,827–11,357	5,548–22,777
ACO savings ($ per beneficiary)	101 (680)	13	−252 to 394	−3,136 to 2,586

distributions. Each mixture component corresponds to an ACO benchmark group (e.g., low or high benchmark), which allows us to capture the dependence of ACO type on benchmark in a semiparametric manner. Although our approach can be extended to alternate choices for the type distribution, we choose the exponential distribution on the basis of the observation that most ACOs generate minimal savings, which suggests that ACO types are concentrated at lower values of the type parameter. Moreover, use of the exponential distribution requires us to estimate only a single parameter for each mixture component, which limits the complexity of the model and prevents overfitting. Let m be the number of mixture components, indexed by j. Then let $\mathcal{M}_1, \ldots, \mathcal{M}_m$ be a disjoint partitioning of the positive real line, where each interval $\mathcal{M}_j$ represents a set of ACO benchmarks. With a slight abuse of notation, let $f(\theta|\lambda_j)$ denote the exponential distribution with parameter λ_j. We assume that the ith ACO's type is drawn from $f(\theta|\lambda_j)$ if the ACO's benchmark μ_i belongs to $\mathcal{M}_j$. For the shock distribution, we assume that the variable ξ is distributed according to the zero-mean Laplace distribution $G(\xi|\sigma)$, where σ is the standard deviation. This specification of the shock distribution satisfies Assumption 2. On the basis of this parameterization, the quantities to be estimated are the m shape parameters $\lambda_1, \ldots, \lambda_m$ and the shock parameter σ. Note that we are not required to determine which mixture component to assign each observation to because the assignment depends only on the benchmark μ_i, which is known from the data. Last, for the investment function, we set $c(x, \theta) = x^2/\theta$, which satisfies Assumption 1.

6.3. Model Idenfiability and Maximum Likelihood Estimation

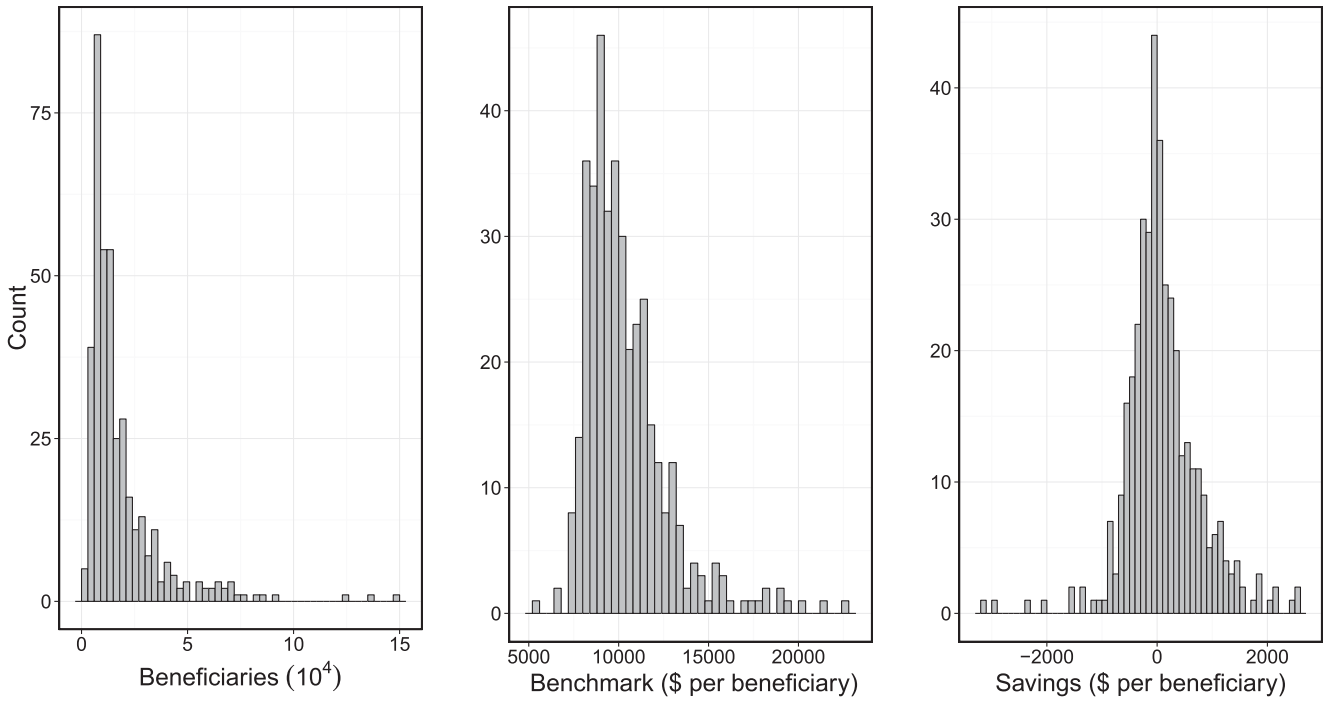
The parameters $(\lambda_1, \ldots, \lambda_m, \sigma)$ can show to be statistically identifiable under an additional condition that is relatively mild. This condition is formalized in Proposition 4.

Proposition 4 (Identification). *If the condition*

$$\int_{\Theta} e^{-\sqrt{2}x(\theta)/\sigma} \left(f(\theta|\lambda) - f(\theta|\tilde{\lambda}) \right) \neq 0 \tag{13}$$

holds for all $\sigma \in \Sigma$ and all $\lambda, \tilde{\lambda} \in \Lambda$ such that $\lambda \neq \tilde{\lambda}$, then the model parameters $(\lambda_1, \ldots, \lambda_m, \sigma)$ are identifiable.

Figure 2. Distribution of Number of Medicare Beneficiaries, Spending Benchmarks, and Savings for 392 ACOs Participating in the MSSP in 2015



Corollary 1. *If Λ and Σ are finite sets and $\bar{\theta}$ is sufficiently large, then the identification condition (13) holds.*

The identification condition (13) ensures that there cannot exist multiple values of the shape parameter λ that give rise to the same savings distribution for a given benchmark group. Identifiability of the model can be guaranteed by placing appropriate assumptions on Λ, Σ, and Θ. Corollary 1 provides one such set of assumptions. The conditions in Corollary 1 are mild because one can construct finite sets Λ and Σ by discretizing a continuous parameter space to an arbitrarily fine degree. Additionally, assuming that $\bar{\theta}$ is very large does not materially change the model or estimation approach because it simply expands the space of possible ACO types.

We now outline a maximum likelihood estimation (MLE) approach for the parameters $(\lambda_1, \ldots, \lambda_m, \sigma)$. Obtaining maximum likelihood estimates for our principal–agent model requires us to solve an *inverse optimization* problem, which refers to the estimation of optimization model parameters from (potentially noisy) solution data (Ahuja and Orlin 2001, Bertsimas et al. 2015, Aswani et al. 2018). The MLE problem takes the form of an inverse optimization problem because the observed savings data represent noisy observations of the ACO's optimal decision (recall that $y = x + \xi$). We first require the following assumption.

Assumption 5. *The data $(\mu_i, b_i, y_i, \theta_i)$, $i = 1, \ldots, n$, are drawn independently from a common distribution.*

Assumption 5 states that the benchmark, number of beneficiaries, type, and savings of an ACO are independent of other ACOs. This assumption is reasonable given that each ACO operates independently and is necessary for tractability of the estimator. Note that Assumption 5 permits dependence between the attributes of a given ACO. We now formalize the estimation problem. Let $\Lambda_1, \ldots, \Lambda_m$ and Σ denote the parameter sets for $\lambda_1, \ldots, \lambda_m$ and σ, respectively.

Proposition 5 (MLE). *The maximum likelihood estimate of $(\lambda_1, \ldots, \lambda_m, \sigma)$ is given by*

$$(\hat{\lambda}, \hat{\sigma}) = \operatorname*{argmax}_{\lambda \in \Lambda, \sigma \in \Sigma} \sum_{i=1}^{n} \log\left(\int_{\Theta} g(y_i - x(\theta) | \sigma, \theta)\, f(\theta | \lambda(\mu_i)) d\theta \right),$$

$$x(\theta) = \operatorname*{argmax}_{x \geq 0} \int_{-\infty}^{\infty} r_+(y, \alpha) \omega(y|x) dy - \gamma x - c(x, \theta),$$

where $\lambda(\mu_i) = \lambda_j$ if $\mu_i \in \mathcal{M}_j$.

In Proposition 5, $x(\theta)$ refers to the optimal savings of a type θ under the contract that generated the data. Because our data set contains the performance of ACOs under the one-sided contract, we use the one-sided shared savings function r_+ (given in (1)) in the ACO's problem in Proposition 5. If the data set represented ACO performance under the two-sided contract, then we would simply replace r_+ with the two-sided payment function r given in (2).[8]

Maximum likelihood estimators are often difficult to solve to global optimality owing to the likelihood function being nonconvex. As a result, estimation approaches that find local maxima of the likelihood function (e.g., expectation maximization) are typically used (Hastie et al. 2005). However, because the number of parameters is relatively small in our setting (assuming that m is not too large), we obtain an approximately optimal solution to the MLE problem as follows. Let Λ, Σ, and Θ be discrete sets. First, we numerically solve the ACO's problem to obtain $x(\theta)$ for each $(t, \sigma) \in \Theta \times \Sigma$ (note that the ACO's problem depends on σ as well as θ). Then, for each $\sigma \in \Sigma$, we evaluate the likelihood function given in Proposition 5 for each $(\lambda_1, \ldots, \lambda_m) \in \Lambda_1 \times \cdots \times \Lambda_m$ and select the parameter vector $(\lambda_1, \ldots, \lambda_m)$ with the largest likelihood.

6.4. Specification of Parameters

6.4.1. Mixture Distribution. If the ACO type distribution is defined by a large number of mixture components, then the resulting model may overfit to the training data and thus not be generalizable outside the sample. Thus, two additional modeling decisions that remain are the number of benchmark groups (i.e., mixture components) and the range of ACO benchmarks covered by each group. We tuned these additional parameters by using k-means clustering to identify the endpoints of the intervals corresponding to each of k benchmark groups and by using 10-fold cross-validation (on the training data) to identify the optimal number of benchmark groups. For the cross-validation step, we fit the model specified above using MLE (described in Section 6.3) and evaluated model performance by computing the likelihood function for the validation set. We performed the cross-validation by varying the number of benchmark groups from one to six and found the optimal number of mixture components to be three. The associated benchmark groups implied by the clustering were $\mathcal{M}_1 = [0, 1.03 \times 10^4]$, $\mathcal{M}_2 = (1.03 \times 10^4, 1.43 \times 10^4]$, and $\mathcal{M}_3 = (1.43 \times 10^4, \infty)$.

6.4.2. ACO Profit Function. To solve the MLE problem, we must also specify the parameters in the ACO's profit function $u(x, \alpha, \beta, \theta)$. We set $\alpha = 0.46$ to represent the "effective" shared savings rate, which we observed from the ACO data to be the average shared savings rate that ACOs received after quality adjustments (a maximum rate of 0.5 multiplied by the average quality score of 92%). Although the quality scores vary from one ACO to the next, we assume a fixed score of 92%

across all ACOs to reduce model complexity during estimation. We find that we obtain a strong model fit despite this simplifying assumption. We explicitly incorporate quality scores in our estimate of ACO performance under the optimal contract by sampling each ACO's quality in the bootstrapping procedure, where it is used to adjust the nominal shared savings rate.

The remaining parameters to be defined are the payment cap C_u, the penalty cap C_ℓ, the minimum savings and losses threshold h, and the ACO's Medicare profit margin γ. We set $C_u = 2{,}000$ and $C_\ell = -1{,}500$ according to MSSP guidelines (Federal Register 2011). We set $h = 200$ because the minimum savings and losses threshold is mandated to be 2% of the ACO's benchmark, and the average benchmark in the data set was $\$10,403$ per beneficiary.[9] Last, we set the ACO's Medicare profit margin as $\gamma = 0.03$ based on a recent report by the Medicare Payment Advisory Commission MedPAC (2010).

6.5. Estimation of Contract Performance

Using the parameter estimates obtained from Section 6, we solve the four optimal contracting problems OC-I, OC-II, OC-III, and OC-IV described in Sections 4 and 5. The space of possible shared savings and subsidy rates is given by $\mathcal{A} = \{0, 0.05, 0.1, \ldots, 0.9\}$ and $\mathcal{B} = \{0, 0.05, 0.1, \ldots, 0.9\}$, respectively. We let the type space be discrete; that is, we set $\Theta = \{1{,}500, 1000, \ldots, 5000\}$, which allows us to formulate and solve OC-I, OC-II, OC-III, and OC-IV as integer optimization models (see Section EC.2 of the online supplemental material). For problems OC-I and OC-III, which include participation constraints, $\bar{u}(\theta)$ is set to the expected ACO payoff under the existing two-sided contract for a type θ ACO. This restricts the set of feasible contracts in problems OC-I and OC-III to those that produce at least as high a payoff for the ACO as the current MSSP contract. For the nonparametric contract, we use the same Θ as above. For each $\theta \in \Theta$, Medicare selects $\rho(\theta)$ from the set $\mathcal{R} = \{0, 20, 40, \ldots, 2000\}$ and $x(\theta)$ from the set $\mathcal{X} = \{0, 20, 40, \ldots, 2000\}$.[10]

Using the estimated model parameters, we simulate ACO performance under the existing MSSP contract and the proposed contracts using a standard nonparametric bootstrapping procedure [see Efron and Tibshirani (1994) for a comprehensive overview of bootstrapping techniques]. In particular, for each ACO, we sample its benchmark spending, actual spending, quality score, and number of beneficiaries, which allows us to estimate the total savings under each contract. We also compute 95% confidence intervals for the savings under each contract and perform bootstrap hypothesis tests to assess the statistical significance of the estimated improvement of the proposed models over the baseline contract.

7. Results and Policy Implications

For validation purposes, we estimate the model parameters using data from a random subset of 275 ACOs (70% of the data) and validate them against an out-of-sample data set containing the performance of the remaining 117 ACOs (30% of the data). Table 3 presents the parameter estimates for each of the three benchmark groups, where $\hat{\lambda}$ is the estimated shape parameter of the exponential distribution and $\hat{\sigma}$ is the estimated standard deviation of the random shock. Note that $\hat{\sigma}$ is the same for all three benchmark groups because we estimate only a single shock distribution to avoid overfitting. The estimated mean of the exponential distributions is given by $1/\hat{\lambda}$, which we may interpret as the average ACO type for the given benchmark cluster. Because the parameter estimates in Table 3 are difficult to interpret directly, we also report the expected optimal savings $\mathbb{E}_t[x(\theta)]$ under the MSSP's current one-sided contract.

To validate our model, we simulated the savings that would be generated under the current contract using the parameter estimates given in Table 3. Figure 3 shows the empirical savings distribution for both the in-sample and out-of-sample data sets and the simulated cumulative distribution function (CDF) from our fitted model (with 10,000 samples). We performed a Kolmogorov–Smirnov (K-S) test to assess the goodness of fit of our model with respect to the empirical savings data (Massey 1951). The K-S statistic (i.e., the maximum vertical distance between the empirical and simulated CDFs) on the out-of-sample data set was 0.076, with an associated p-value of $p > 0.2$. This p-value implies a *failure* to reject the null hypothesis that the two CDFs were generated under different statistical models at a confidence level of 80%, suggesting a strong model fit.

Our results show that, in general, ACOs with high benchmarks are more likely to be high type and thus more likely to generate positive savings. As shown in Table 3, we estimated the expected per-beneficiary savings of low ($< \$10,300$), intermediate ($\$10,300 - \$14,300$), and high ($> \$14,300$) benchmark ACOs to be \$1, \$140, and \$260, respectively. Figure 4 provides an accompanying visualization that shows the estimated type distribution for each benchmark group. Observe that the type distribution for low-benchmark ACOs is concentrated near $\theta = 0$, whereas the distributions for the intermediate- and high-benchmark groups place mass across a range of values of θ. As one might expect, the parameter estimate $\hat{\lambda}_j$ for the intermediate-benchmark group falls between the low- and high-benchmark values. We numerically found the type threshold θ_s discussed in Section 4.2 to be approximately 80 for our choice of the ACO investment function. As can be seen from Figure 4, ACOs in the intermediate- and high-benchmark groups have a high probability of being above this type threshold, suggesting that the subsidy may be useful in

Table 3. Maximum Likelihood Parameter Estimates for Three Benchmark Groups

Benchmark	No. of ACOs	$\hat{\lambda}$	$\mathbb{E}[\theta]$	$\hat{\sigma}$	$\mathbb{E}[x(\theta)]$
<$10,300	228	1	1	550	$1
$10,300–$14,300	140	1.1×10^{-3}	900	550	$140
>$14,300	24	6.7×10^{-4}	1,500	550	$260

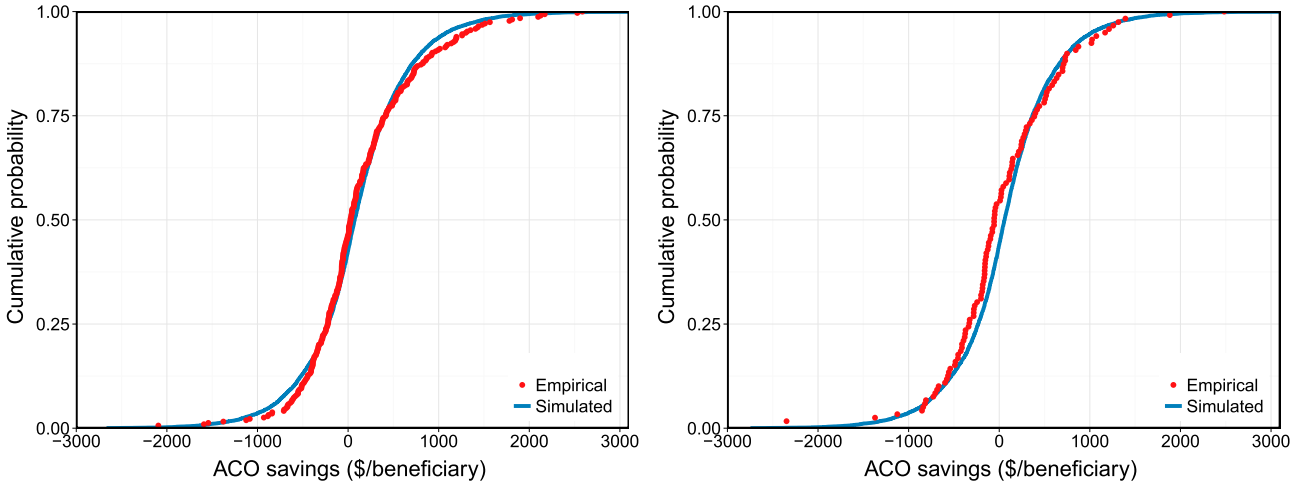
boosting savings from these groups. In contrast, because low-benchmark ACOs are low type, the subsidy may not be effective in improving savings from that group.

The observation that high-benchmark ACOs tend to save more may be a consequence of the benchmark calculation methodology that is currently used by the MSSP. As discussed previously, an ACO's benchmark is largely determined by its historical spending. Because there are multiple factors that contribute to an ACO's historical spending, such as regional variations in healthcare costs, it is difficult to pinpoint exactly why a particular ACO has high historical spending. However, our results suggest that ACOs with high benchmark expenditures may have an easier time reducing spending because they were historically cost inefficient and thus have "more room" for improvement. By contrast, an ACO with a low benchmark may be less able to generate savings because it has historically been cost efficient, and thus additional reductions in spending are more difficult to attain.

Figure 5 shows the optimal contract parameters for OC-I and OC-II. Note that the right panel of Figure 5 exhibits the behavior predicted by Proposition 1, where low-type ACOs are assigned a shared savings rate where $\alpha(\theta) \geq 1/2$. Additionally, the results shown in Figure 5 align with Proposition 2, where high-type ACOs are assigned relatively higher subsidy rates and lower shared savings rates, with the converse holding true for low-type ACOs. Table 4 shows the estimated ACO payoff, Medicare savings, and total welfare under four contracts: the subsidy-based contract including the participation constraint (OC-I), the subsidy-based contract with endogenous participation (OC-II), the nonparametric contract with participation constraint (OC-III), and the nonparametric contract with endogenous participation (OC-IV). For comparison purposes, we simulate the performance of the 392 ACOs once they enter the current two-sided contract of the MSSP and report the results in the "Baseline" column. We also report the difference between the baseline and optimal contracts, denoted by Δ, as well as a one-sided p-value to test the significance of the difference [see Efron and Tibshirani (1994) for an overview of hypothesis testing with the bootstrap]. We also report the 95% confidence intervals for the payoff estimates under each contract and the performance gap Δ. We note here that although we held quality constant throughout our analysis thus far, for simulation purposes, we also sample quality scores for each ACO in the bootstrap and calculate the savings according to the quality-adjusted value of the shared savings rate α.

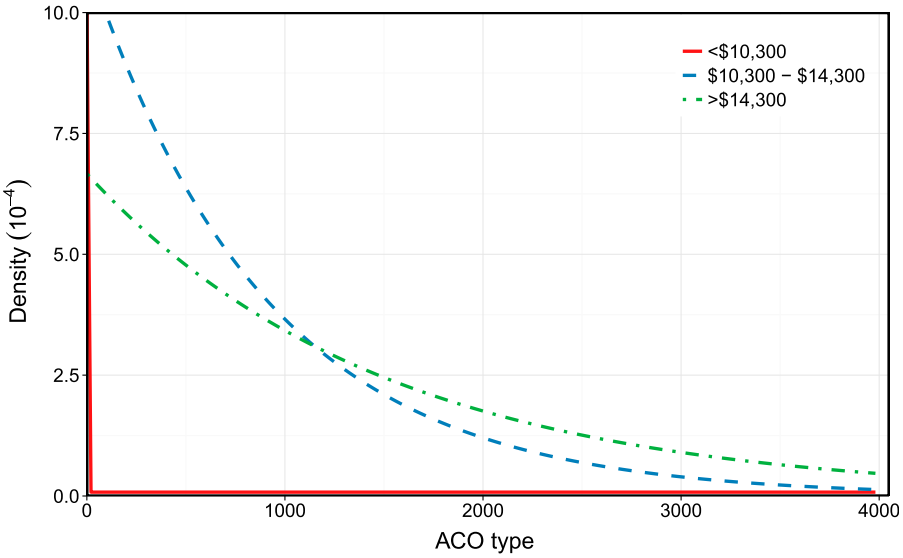
In the subsidy-based contract (OC-I), the ACOs and Medicare both experience an improvement compared with the baseline contract. This finding aligns with the result in Proposition 3. Specifically, ACO payoff increases by 12% ($282 to $316 million), and Medicare savings increase by 43% ($146 to $207 million), relative to their baseline levels. Note also that the increase in total welfare (the sum of ACO payoff and Medicare savings) is 22% under the subsidy-based contract. As an intermediate step in our analysis, we also solved for the optimal non-subsidy-based contract where β is held

Figure 3. (Color online) Empirical and Simulated Savings Under One-Sided Contract for In-Sample and Out-of-Sample Data



In-sample. Out-of-sample.

Figure 4. (Color online) Estimated ACO Type Distributions for Low (<$10,300), Intermediate ($10,300–$14,300), and High (>$14,300) Benchmark Groups



fixed at 0 and the contracting problem is solved over $\alpha(\theta)$ only to identify the optimal set of shared savings rates. The improvement in Medicare savings from optimizing over the shared savings rate alone was found to be negligible, so those results are omitted. This result suggests that in the absence of a subsidy, Medicare cannot boost savings by adjusting the shared savings rate *without* compromising ACO payoff and, by extension, their participation in the MSSP. Although it has been suggested that the MSSP should increase the shared savings rate to strengthen incentives for ACOs (Chernew et al. 2014), our analysis suggests that doing so will only further reduce Medicare savings owing to the associated increase in ACO payments.

In Section 4.2, we showed that subsidizing the investments made by ACOs to reduce healthcare spending can improve both Medicare savings and ACO payoff. This result is validated by our empirical analysis and the findings in the Table 4 under OC-I. Further, the vast majority of ACOs in the MSSP remain under the initial one-sided contract and are therefore not penalized for accruing losses. However, of the 392 ACOs in our data set, 28 attained quality scores that would result in an effective shared savings rate that is less than 0.5 in the two-sided contract. In light of Proposition 1, these ACOs are particularly at risk for dropping out of the MSSP once they transition to the two-sided contract. Our results suggest that subsidizing ACO investments may be useful in mitigating this risk.

Figure 5. (Color online) Optimal Contract Parameters for Subsidy-Based Contract with Participation Constraint (OC-I) and with Endogenous Participation (OC-II)

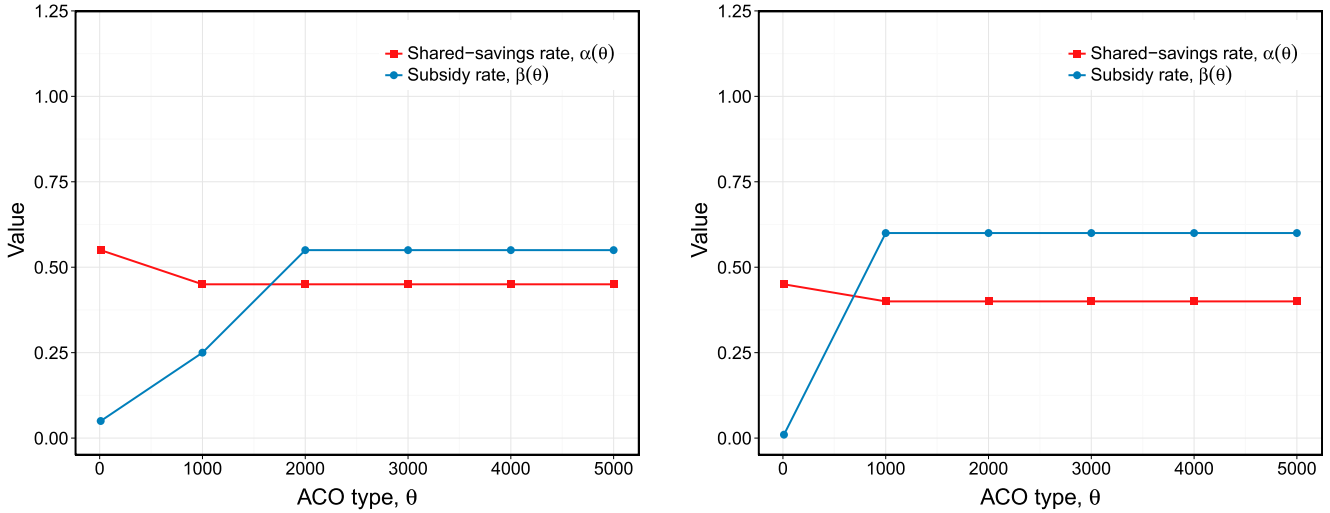


Table 4. Bootstrap Estimates for Subsidy-Based Contract with Participation Constraint (OC-I), Subsidy-Based Contract with Endogenous Participation (OC-II), Nonparametric Contract with Participation Constraint (OC-III), and Nonparametric Contract with Endogenous Participation (OC-IV), in Millions

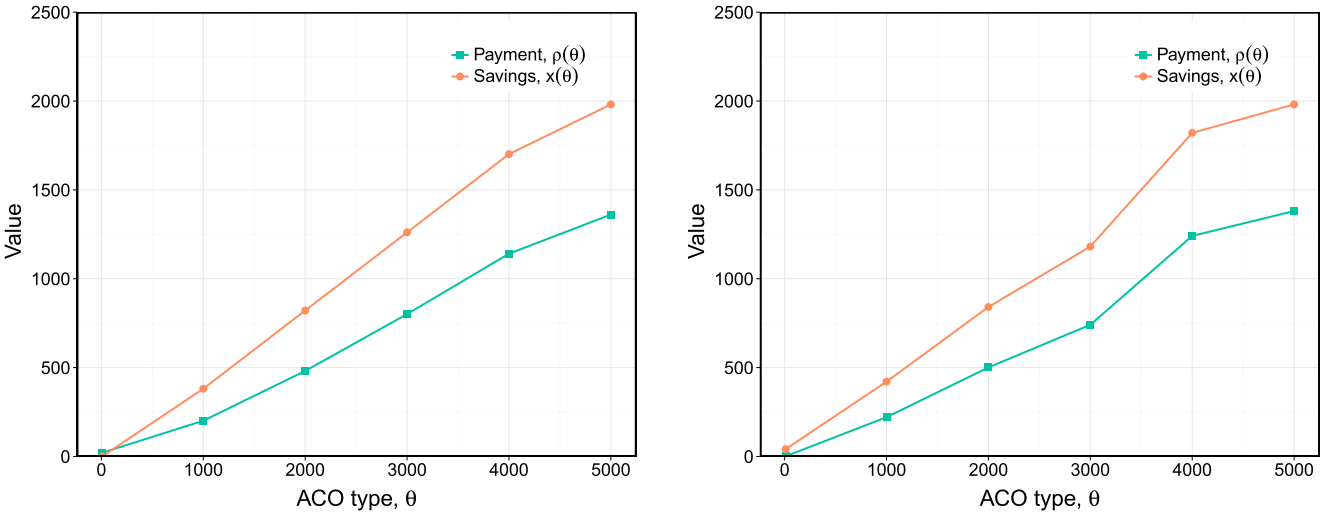
	Baseline		Optimal		Δ		
Model	Mean	95% Confidence interval	Mean	95% Confidence interval	Mean	95% Confidence interval	*p*-Value
OC-I							
ACOs	$282	($188, $382)	$316	($216, $423)	$34	($22, $49)	<0.01
Medicare	$146	($39, $260)	$207	($97, $328)	$62	($48, $79)	<0.01
Total	$427	($234, $642)	$523	($320, $741)	$96	($78, $118)	<0.01
OC-II							
ACOs	$282	($188, $382)	$126	($87, $174)	–$156	(–$212, –$97)	<0.01
Medicare	$146	($39, $260)	$377	($295, $455)	$231	($164, $304)	<0.01
Total	$427	($234, $642)	$503	($383, $614)	$72	(–$48, $189)	>0.1
OC-III							
ACOs	$282	($188, $382)	$301	($265, $346)	$19	(–$73, $109)	>0.1
Medicare	$146	($39, $260)	$237	($92, $381)	$91	($30, $152)	<0.01
Total	$427	($234, $642)	$538	($325, $745)	$110	($82, $122)	<0.01
OC-IV							
ACOs	$282	($188, $382)	$120	($95, $146)	–$164	(–$222, –$94)	<0.01
Medicare	$146	($39, $260)	$384	($283, $490)	$236	($163, $313)	<0.01
Total	$427	($234, $642)	$504	($383, $615)	$72	(–$48, $189)	<0.01

The second block in Table 4 (OC-II) shows the estimated performance of the subsidy-based contract when ACO participation is determined endogenously instead of being enforced as a constraint in the optimal contracting problem. As we would expect, relaxing the participating constraint increases Medicare's savings significantly. However, unlike the win–win scenario that arises under OC-I, many ACOs will see a decrease in payments under OC-II, which would likely lead to some ACOs dropping out of the MSSP. Although Medicare should expect total savings to increase under this contract, abandonment of the MSSP by a large number of ACOs may make it more challenging for Medicare to pursue a nationwide transition away from fee-for-service and toward voluntary pay-for-performance programs. Indeed, maintaining a critical mass of ACOs is believed to be important to the success of the MSSP (Rosenthal et al. 2011). Therefore, the CMS should carefully weigh the possible trade-offs between total savings and ACO participation before making any adjustments to the MSSP contract.

The third block in Table 4 (OC-III) shows the estimated performance of the optimal nonparametric contract when the participation constraint is included. The optimal contract parameters are shown in Figure 6. Note that Figure 6 is expressed in dollars per beneficiary (e.g., for the median ACO with a benchmark $9,863 per beneficiary, a value of 100 corresponds to approximately 1% of the ACO's benchmark). Because of its generality, this contract represents the best possible performance of any MSSP contract given the estimated type distribution. Under the nonparametric contract, Medicare's total savings increases from $146 million to $237 million, which represents a 63% increase in total savings compared with the baseline contract. The nonparametric contract also improves on the subsidy-based contract ($237 million vs. $207 million). This is unsurprising given the additional flexibility provided by the nonparametric contract. However, it is worth noting that a significant share of the improvement potential associated with the nonparametric contract ($91 over baseline) is captured by the subsidy-based contract ($62 million over baseline). This result suggests that introducing a straightforward investment subsidy to the MSSP carries much of the benefit associated with completely redesigning the MSSP contract. The fourth block in Table 4 shows the estimated performance when ACO participation is determined endogenously. Unsurprisingly, this contract generates the highest savings overall for Medicare of the four models.

In general, Table 4 suggests that a total overhaul of the MSSP contract, as represented by the nonparametric contracting problems OC-III and OC-IV, is likely to generate higher savings than the the subsidy-based contract. However, the sheer scale of the MSSP ($70+ billion in Medicare spending) suggests that there exists significant institutional inertia behind the current program, making a total overhaul of the program potentially difficult for Medicare. In contrast, although the subsidy-based contract generates lower savings, it is achieved through a less dramatic modification of the current contract, making it potentially more practical to implement. This trade-off between savings and implementability should be considered as the CMS decides how to move forward with the MSSP.

Figure 6. (Color online) Optimal Contract Parameters for Nonparametric Contract with Participation Constraint (OC-III) and with Endogenous Participation (OC-IV), in Dollars per Beneficiary



OC-III. OC-IV .

8. Conclusion

The MSSP was created to incentivize Medicare providers to form ACOs and reduce spending on healthcare delivery. To date, results from the MSSP have been mixed, with one-third of participating ACOs failing to generate sufficient savings to receive reward payments from Medicare. As a consequence, total Medicare savings has been modest, and the MSSP currently faces the risk of many ACOs abandoning the program entirely. In this paper, we proposed a new type of contract to address the challenges in the MSSP. In addition to the shared savings payment, we proposed that the MSSP incorporate direct subsidies to partially reimburse ACO investments toward efficiency improvements. We showed that selecting the shared savings and subsidy rates appropriately yields a contract that dominates all possible contracts within the current, nonsubsidized program, in the sense that it boosts both Medicare savings and ACO payments. We also quantified the improvement potential through structural estimation of the principal–agent model and found that switching to the proposed contract can increase both Medicare savings and ACO payments. We also found that the proposed contract performs well in comparison with a nonparametric contract that represents the theoretical best performance that can be expected from the MSSP under the estimated agent model. We also found that ACOs with low benchmarks face difficulties in generating savings. The average savings of an ACO with a benchmark greater than \$14,300 per beneficiary was estimated to be \$240 (≈2%), whereas the average savings of an ACO with a benchmark of less than \$10,300 per beneficiary was estimated to be effectively zero.

It is unlikely that there is a single solution to the challenges currently faced by the MSSP. More likely, a multifaceted restructuring of the existing contract will be required to boost Medicare savings and increase the appeal of the MSSP to existing and prospective ACOs. Our analytical and empirical results suggest that ACO investment subsidies have the potential to play an important role in this restructuring. Moreover, as the MSSP continues and more data are collected, a clearer picture of ACO performance within the MSSP will emerge. We hope that the empirical approach developed in this paper can serve as a springboard for future analyses of the MSSP.

Acknowledgments

The authors thank Area Editor Chung Piaw Teo, the associate editor, and two referees for thoughtful feedback, which has considerably improved this paper.

Appendix. Proofs

For the purposes of this section, let $R(x,\alpha) = \mathbb{E}[r(y,\alpha)] = \int_{-\infty}^{\infty} r(y,\alpha)\omega(y|x)dy$ be the expected shared savings payment, and let $S(x,\beta) = \mathbb{E}[s(y,\beta)] = \int_{-\infty}^{\infty} s(y,\beta)\omega(y|x)dy$ be the expected subsidy payment. We first present a lemma that is useful in the proofs to follow.

Lemma 4. *The derivative of the expected shared savings payment with respect to x is*

$$(\partial/\partial x)R(x,\alpha) = (1-\alpha)(hg(-h-x) + G(-h-x) - G(C_\ell - x)) + \alpha(hg(h-x) + G(C_u - x) - G(h-x)).$$

The derivative of the expected subsidy payment with respect to x is

$$(\partial/\partial x)S(x,\beta) = \beta((\partial/\partial x)c(x,\theta)(1-G(h-x)) + c(x,\theta)g(h-x)).$$

Proof. Consider the shared savings term first. Integrating over the piecewise expression given in (2), we have

$$\begin{aligned}R(x,\alpha) &= \int_{-\infty}^{\infty} r(y,\alpha)\omega(y|x)dy \\ &= (1-\alpha)C_\ell \int_{-\infty}^{C_\ell} \omega(y|x)dy + (1-\alpha)\int_{C_\ell}^{-h} y\omega(y|x)dy \\ &\quad + \alpha\int_{h}^{C_u} y\omega(y|x)dy + \alpha C_u \int_{C_u}^{\infty}\omega(y|x)dy.\end{aligned}$$

Because $y = x+\xi$, we may apply a change of variables and write the expression above equivalently as

$$\begin{aligned}R(x,\alpha) &= (1-\alpha)C_\ell \int_{-\infty}^{C_\ell - x} g(\xi)d\xi + (1-\alpha)\int_{C_\ell - x}^{-h-x}(x+\xi)g(\xi)d\xi \\ &\quad + \alpha\int_{h-x}^{C_u-x}(x+\xi)g(\xi)d\xi + \alpha C_u\int_{C_u-x}^{\infty} g(\xi)d\xi.\end{aligned}$$

Applying Leibniz's rule to each of the four terms above yields

$$\begin{aligned}(\partial/\partial x)\partial R(x,\alpha) &= -(1-\alpha)C_\ell g(C_\ell - x) + (1-\alpha)(hg(-h-x) \\ &\quad + C_\ell g(C_\ell - x) + G(-h-x) - G(C_\ell - x)) \\ &\quad + \alpha(-C_u g(C_u - x) + hg(h-x) + G(C_u - x) \\ &\quad - G(h-x)) + \alpha C_u g(C_u - x) \\ &= (1-\alpha)(hg(-h-x) + G(-h-x) - G(C_\ell - x)) \\ &\quad + \alpha(hg(h-x) + G(C_u - x) - G(h-x)),\end{aligned}$$

where the second line is obtained by canceling out terms. For the subsidy term, we integrate over the expression given in (3) to obtain $S(x,\beta) = \beta c(x,\theta)\int_h^\infty \omega(y|x)dy = \beta c(x,\theta)\int_{h-x}^{\infty} g(\xi)d\xi$. Applying the chain rule to this expression yields

$$\begin{aligned}(\partial/\partial x)S(x,\beta) &= \beta\Big((\partial/\partial x)c(x,\theta)(1-G(h-x)) \\ &\quad + c(x,\theta)(\partial/\partial x)\Big(\int_{h-x}^{\infty} g(\xi)d\xi\Big)\Big) \\ &= \beta((\partial/\partial x)c(x,\theta)(1-G(h-x)) \\ &\quad + c(x,\theta)g(h-x)). \quad \square\end{aligned}$$

Proof of Lemma 1. To show that $u(x,\alpha,\beta,\theta)$ is strictly concave over $[0,\bar{x}]$ for any θ, it suffices to show that $(\partial/\partial x)u(x,\alpha,\beta,\theta) < 0$ for all $x \in [0,\bar{x}]$ and $\theta \in \Theta$. Pick some x and θ such that $(\partial^2/\partial x^2)u(x)$ exists. (Note that $(\partial^2/\partial x^2)u(x)$ may be undefined for certain values of x owing to the possible nondifferentiability of the density $g(\xi)$ at some points. In these cases, we may consider the left and right derivatives of $u(x,\alpha,\beta,\theta)$ and apply a similar argument.) By twice differentiating the expression for $u(x,\alpha,\beta,\theta)$ given in (4), we have $(\partial^2/\partial x^2)u(x) = (\partial^2/\partial x^2)R(x,\alpha) + (\partial^2/\partial x^2)S(x,\beta) - (\partial^2/\partial x^2)c(x,\theta)$. We consider each of the three terms in this equation in sequence. For the shared savings term $(\partial^2/\partial x^2)R(x,\alpha)$, differentiating the expression for $(\partial/\partial x)R(x,\alpha)$ given in Lemma 4 yields $(\partial^2/\partial x^2)R(x,\alpha) = (1-\alpha)(-hg'(-h-x) - g(-h-x) + g(C_\ell - x)) + \alpha(-hg'(h-x) - g(C_u - x) + g(h-x))$. Because g is increasing over $(-\infty,0)$ and $C_\ell \le -h$, we have $g(C_\ell - x) \le g(-h-x)$ for all x. Further, we have $g'(-h-x) \ge 0$ and thus $-hg'(-h-x) \le 0$ for all x. It follows that $(1-\alpha)(-hg'(-h-x) - g(-h-x) + g(C_\ell - x)) \le 0$. Hence, we may drop the first term in the expression for $(\partial^2/\partial x^2)R(x,\alpha)$ above to obtain $(\partial^2/\partial x^2)R(x,\alpha) \le \alpha(-hg'\cdot(h-x) - g(C_u - x) + g(h-x))$. Now, by dropping the negative term $-g(C_u - x)$ and noting that $-hg'(h-x) \le h\bar{g}'$ and $g(h-x) \le \bar{g}$, we have $(\partial^2/\partial x^2)R(x,\alpha) \le \bar{\alpha}(h\bar{g}' + \bar{g})$. For the subsidy term $(\partial^2/\partial x^2)S(x,\beta)$, we differentiate the expression for $(\partial/\partial x)S(x,\beta)$ given in Lemma 4 to obtain

$$\begin{aligned}(\partial^2/\partial x^2)S(x,\beta) &= \beta((1-G(h-x))(\partial^2/\partial x^2)c(x,\theta) \\ &\quad + 2g(h-x)(\partial/\partial x)c(x,\theta) - c(x,\theta)g'(h-x)), \\ &\le \bar{\beta}((\partial^2/\partial x^2)c(x,\theta) + 2\bar{g}(\partial/\partial x)c(x,\theta)),\end{aligned}$$

where the inequality follows by noting that $1 - G(h-x) \le 1$ and $g(h-x) \le \bar{g}$. Combining the bounds for $(\partial^2/\partial x^2)R(x,\alpha)$ and $(\partial^2/\partial x^2)S(x,\beta)$, we can now write

$$\begin{aligned}(\partial^2/\partial x^2)u(x) &\le \bar{\alpha}(h\bar{g}' + \bar{g}) + \bar{\beta}((\partial^2/\partial x^2)c(x,\theta) \\ &\quad + 2\bar{g}(\partial/\partial x)c(x,\theta)) - (\partial^2/\partial x^2)c(x,\theta) \\ &= \bar{\beta}(h\bar{g}' + \bar{g}) + 2\beta\bar{g}(\partial/\partial x)c(x,\theta) \\ &\quad - (1-\bar{\beta})(\partial^2/\partial x^2)c(x,\theta) \\ &< 0,\end{aligned}$$

where the final strict inequality follows from Assumption 3. $\square$

Proof of Lemma 2. Let α and β be fixed. We first prove that $x(\theta)$ is strictly increasing over $[\theta_{\alpha,\beta},\bar{\theta}]$. We wish to show that $(d/d\theta)x(\theta) > 0$ for any $\theta \in [\theta_{\alpha,\beta},\bar{\theta}]$. Pick any such θ. Because $x(\theta) > 0$, it follows that $x(\theta)$ is a solution to the first-order condition $(\partial/\partial x)u(x,\alpha,\beta,\theta) = (\partial/\partial x)R(x,\alpha) + (\partial/\partial x)S(x,\beta) - \gamma - (\partial/\partial x)c(x,\theta) = 0$. By the implicit function theorem, we have $(d/d\theta)x(\theta) = -(\partial^2/\partial x\partial\theta)u(x,\alpha,\beta,\theta)/(\partial^2/\partial x^2)u(x,\alpha,\beta,\theta)$. Because $R(x,\alpha)$ and $S(x,\beta)$ do not depend on θ, we have $-(\partial^2/\partial x\partial\theta)u(x) = (\partial^2/\partial x\partial\theta)c(x,\theta) < 0$, where the inequality follows from part (iii) of Assumption 1. Further, by Proposition 1, we also have $(\partial^2/\partial x^2)u(x,\alpha,\beta,\theta) < 0$, and thus $(d/d\theta)x(\theta) > 0$ over $[\theta_{\alpha,\beta},\bar{\theta}]$. It remains to show that x is nondecreasing over $[\underline{\theta},\theta_{\alpha,\beta}]$, which follows immediately from the definition of $\theta_{\alpha,\beta}$ and the observation that x is bounded below by 0. $\square$

Proof of Lemma 3. This proof proceeds in a manner similar to Lemma 2. Pick some $\alpha \in \mathcal{A}$ and $\theta \in [\theta_{\alpha,0},\bar{\theta}]$. Now pick any $\beta \in [\underline{\beta},\bar{\beta}]$. Because $\theta \ge \theta_{\alpha,0}$, we have $x(\beta) > 0$. It follows that $x(\beta)$ is a solution to the first-order condition $(\partial/\partial x)R(x,\alpha) + (\partial/\partial x)S(x,\beta) - \gamma - (\partial/\partial x)c(x,\theta) = 0$. By the implicit function theorem, we have $(d/d\beta)x(\theta) = -(\partial^2/\partial x\partial\beta)u(x,\alpha,\beta,\theta)/(\partial^2/\partial x^2)u(x)$. Because $R(x,\alpha)$ and $c(x,\theta)$ do not depend on β, we have $-(\partial^2/\partial x\partial\beta)u(x) = -(\partial^2/\partial x\partial\beta)S(x,\beta) = -(\partial/\partial x)c(x,\theta)(1-G(h-x)) + c(x,\theta)g(h-x) < 0$. Further, by Lemma 1, we have $(\partial^2/\partial x^2)u(x,\alpha,\beta,\theta) < 0$. It follows that $(d/d\beta)x(\theta) > 0$. To show that $u(x(\theta))$ is increasing in β, note that $S(x,\beta)$ is strictly increasing in β, and therefore so is the ACO's payoff $u(x)$ for any x. It follows that the ACO's optimal payoff $u(x(\theta))$ must be strictly increasing in β as well. $\square$

Proof of Proposition 1. We first show that $\lim_{\theta\to 0} x(\theta) = 0$ for any α. Pick some α, and suppose that for $\epsilon > 0$ we have $x(\theta) > \epsilon$ for all $\theta > 0$. Clearly, the shared savings term $R(x(\theta),\alpha)$ is bounded above by $x(\theta)$. The ACO's optimal payoff can then be bounded above by $u(x(\theta)) \le x(\theta) - \gamma x(\theta) - (1-\beta)c(x(\theta),\theta)$. Because $x(\theta) > \epsilon$ for all $\theta > 0$, and because by Assumption 1 $c(x,\theta)$ is increasing in x, we have $c(\epsilon,\theta) \le c(x(\theta),\theta)$ for all $\theta > 0$. Also by Assumption 1, we have $c(\epsilon,\theta) \longrightarrow \infty$, and thus $\lim_{\theta\to 0} c(x(\theta),\theta) = \infty$. It follows

that $\lim_{\theta\to 0}\ u(x(\theta)) = -\infty$. However, it is straightforward to show that $u(0,\theta)$ is bounded from below, which yields a contradiction. We conclude that $\lim_{\theta\to 0} x(\theta) = 0$. We now show that there exists $\theta_0' > 0$ such that $\beta(\theta) = 0$ is an optimal subsidy rate for $\theta \le \theta_0'$. This follows immediately from the fact that $\lim_{\theta\to 0}\ v(\alpha,0,\theta) = -R(0,\alpha) \ge -R(0,\alpha) - \lim_{\theta\to 0}\ \beta c(x(\theta),\theta) = \lim_{\theta\to 0}\ v(\alpha,\beta,\theta)$ for any $\beta > 0$ and $\alpha > 0$. We now show that there exists θ_0 such that $\alpha(\theta) \ge 1/2$ for all $\theta \le \theta_0$ for sufficiently large C_u and $|C_\ell|$. Using the expression for $R(x,\alpha)$ given in Lemma 4, we have $\lim_{C_u\to\infty}\lim_{C_u\to\infty}\lim_{\theta\to 0} R(x(\theta),\alpha) = (1-\alpha)\int_{-\infty}^{-h}\xi g(\xi)\,d\xi + \alpha\int_h^{\infty}\xi g(\xi)\,d\xi = (1-\alpha)\mathbb{E}[\xi|\xi\le -h] + \alpha\mathbb{E}[\xi|\xi\ge h] = (2\alpha-1)\mathbb{E}[\xi|\xi\ge h]$. Note that $u(x) \le R(x,\alpha)$ for all $x \ge 0$. Therefore, for C_u and $|C_\ell|$ sufficiently large, we have $\lim_{\theta\to 0}\ u(x(\theta)) \le \lim_{\theta\to 0} R(x(\theta),\alpha) < 0$ if $(2\alpha-1) < 0$. Because the participation constraint requires $u(x(\theta)) \ge \bar{u}(\theta) \ge 0$, it follows that $\alpha(\theta) \ge 1/2$ for sufficiently small θ. □

Proof of Proposition 2. Let $\theta_L > 0$ be fixed. We first show that $\beta^*(\theta_H) \ge \beta^*(\theta_L)$ for sufficiently large θ_H. It suffices to show that for any $\alpha \in \mathcal{A}$, $\lim_{\theta\to\infty}(d/d\beta)v(x(\theta),\beta,\theta)|_{\beta=\beta^*(\theta_L)} > 0$. Taking the total derivative of v with respect to β yields $(d/\,d\beta)v(x(\theta),\beta,\theta) = (\partial/\partial x)v(x,\beta,\theta)(d/d\beta)x(\theta) + (\partial/\partial x)\partial v(x,\beta,\theta)$. First, expanding the term $(\partial/\partial x)v(x,\beta,\theta)$ for all $x \in [0,\bar{x}]$, we have

$$\begin{aligned}(\partial/\partial x)v(x,\beta,\theta) &= 1 - (\partial/\partial x)R(x,\alpha) - (\partial/\partial x)S(x,\beta)\\ &= 1 - (\partial/\partial x)R(x,\alpha) - \beta((\partial/\partial x)c(x,\theta)\\ &\quad\cdot(1-G(h-x)) + c(x,\theta)g(h-x))\\ &\ge 1 - (\partial/\partial x)R(x,\alpha) - \beta((\partial/\partial x)c(x,\theta) + c(x,\theta)\bar{g})\\ &\ge \delta - \beta((\partial/\partial x)c(x,\theta) + c(x,\theta)\bar{g}),\end{aligned}$$

where in the final inequality we have $\delta = 1 - (\partial/\partial x)R(x,\alpha) > 0$ for all $x \in [0,\bar{x}]$ and $\alpha \in \mathcal{A}$ by Assumption 4. Because by Assumption 1 the functions $c(x,\theta)$ and $(\partial/\partial x)c(x,\theta)$ are both decreasing in θ, it follows that $\lim_{\theta\to\infty}((\partial/\partial x)c(x,\theta) + c(x,\theta)\bar{g}) = 0$ for all $x \in [0,\bar{x}]$. Hence, $\lim_{\theta\to\infty}(\partial/\partial x)v(x,\beta,\theta)|_{\beta=\beta^*(\theta_L)} > 0$ for all $x \in [0,\bar{x}]$. Further, because by Proposition 3, $(d/d\beta)x(\theta) > 0$ for sufficiently large θ, it follows that $\lim_{\theta\to\infty}(\partial/\partial x)v(x,\beta,\theta)(d/d\beta)x(\theta)|_{\beta=\beta^*(\theta_L)} > 0$. Now note that for the second term, we have $\lim_{\theta\to\infty}(\partial/\partial\beta)v(x,\beta,\theta) = \lim_{\theta\to\infty} c(x,\theta)(1-G(h-x)) \le \lim_{\theta\to\infty} c(\bar{x},\theta) = 0$. It follows that $\lim_{\theta\to\infty}(\partial/\partial\beta)v(x(\beta),\beta,\theta)|_{\beta=\beta^*(\theta_L)} > 0$. Thus, $\beta^*(\theta_H) \ge \beta^*(\theta_L)$. We now show that $\alpha^*(\theta_H) \le \alpha^*(\theta_L)$. Suppose that $\alpha^*(\theta_H) > \alpha^*(\theta_L)$. Because the ACO's payoff $u(x)$ is strictly increasing in α and β, if $\beta^*(\theta_H) \ge \beta^*(\theta_L)$ and $\alpha^*(\theta_H) > \alpha^*(\theta_L)$, then a type θ_L ACO could earn a higher payoff by reporting its type to be θ_H instead of θ_L, which violates incentive compatibility. Therefore, $\alpha^*(\theta_H) \le \alpha^*(\theta_L)$. □

Proof of Proposition 3. We first establish two supporting results. For conciseness, we suppress dependence of $x(\cdot)$ on α and β and dependence of $v(\cdot)$ on α. First, we show that Medicare's savings function $v(x(\theta),\beta,\theta)$ is continuous in β. Note that $v(x,\beta,\theta)$ is continuous in x and β. By the Berge maximum theorem, $x(\theta)$ is upper hemicontinuous in β. Because by Lemma 1 the optimal savings $x(\theta)$ is unique, $x(\theta)$ is also continuous in β. Hence, $v(x(\theta),\beta,\theta)$ is continuous in β. Next, we show that for all $\alpha \in \mathcal{A}$ and $\theta \in \Theta$, there exists $\beta_s > 0$ such that $(\partial/\partial x)\partial v(x,\beta,\theta) > 0$. Pick some α and θ. Now define $\delta = 1 - (\partial/\partial x)R(x,\alpha) > 0$, where the inequality follows from Assumption 4. Following the proof of Proposition 2, we have $(\partial/\partial x)v(x,\beta,\theta) \ge \delta - \beta((\partial/\partial x)c(\bar{x},\theta) + c(\bar{x},\theta)\bar{g})$. Because $\delta > 0$, $(\partial/\partial x)v(x,\beta,\theta) > 0$ for sufficiently small β. We now prove the main result. We wish to show that there exists some $\theta_s > 0$ such that for any $\theta \ge \theta_s$ and $\alpha \in \mathcal{A}$, there exists $\tilde{\beta} > 0$ such that $v(x(\theta),0,\theta) < v(x(\theta),\tilde{\beta},\theta)$. Because $v(x(\theta),\beta,\theta)$ is continuous in β, for any $\tilde{\beta}$ we can apply the mean value theorem to write $v(x(\theta),\tilde{\beta},\theta) = v(x(\theta),0,\theta) + \tilde{\beta}\cdot(d/d\beta)v(x(\theta),\beta,\theta)|_{\beta=\eta\tilde{\beta}}$ for some $\eta \in [0,1]$. Therefore, to show that $v(x(\theta),0,\theta) < v(x(\theta),\tilde{\beta},\theta)$ for each $\theta \ge \theta_s$ and $\tilde{\beta}$, it suffices to show that $\tilde{\beta}(d/d\beta)v(x(\theta),\beta,\theta)|_{\beta=\eta\tilde{\beta}} > 0$ for sufficiently large θ and an associated $\tilde{\beta}$ and $\eta \in [0,1]$. We do so by showing that there exists θ_s such that for all $\theta \ge \theta_s$ there exists a constant $\tilde{\beta}$ such that $(d/d\beta)v(x(\theta),\beta,\theta) > 0$ for all $\beta \in [0,\tilde{\beta}]$ and $\alpha \in \mathcal{A}$. Pick $\tilde{\beta} = \beta_s$. Applying the chain rule yields $(d/d\beta)v(x(\theta),\beta,\theta) = (\partial/\partial x)v(x,\beta,\theta)(d/d\beta)x(\theta) + (\partial/\partial\beta)v(x,\beta,\theta)$. Note that $(\partial/\partial x)v(x,\beta,\theta) > 0$ for all $\alpha \in \mathcal{A}$ if $\beta \le \beta_s$ from the earlier argument, and by Lemma 3, $(d/d\beta)x(\theta) > 0$ for all β if $\theta \ge \theta_{\alpha,0}$. It follows that $(\partial/\partial x)v(x,\beta,\theta)(d/d\beta)x(\theta) > 0$ for any $\alpha \in \mathcal{A}$, $\theta \ge \theta_{\alpha,0}$, and $\beta \in [0,\beta_s]$. We also have $(\partial/\partial\beta)v(x,\beta,\theta) = -c(x,\theta)(1-G(h-x))$, which, by Assumption 1, can be made arbitrarily small by picking θ to be large. It follows that for all $\alpha \in \mathcal{A}$, $\beta \in [0,\beta_s]$, $\lim_{\theta\to\infty}(d/d\beta)v(x(\theta),\beta,\theta) > 0$. Thus, there exists θ_s such that for all $\theta \ge \theta_s$ and $\alpha \in \mathcal{A}$, there exists $\tilde{\beta}$ such that $v(x(\theta),0,\theta) < v(x(\theta),\tilde{\beta},\theta)$. It follows that $V_s^*(\theta) > V_0^*(\theta)$. Because for any $x > 0$, the ACO's payoff $u(x)$ is strictly increasing in β, it follows that $U_s^*(\theta) > U_0^*(\theta)$ as well. □

Proof of Proposition 4. We prove that the model is identifiable for a single benchmark group because the proof extends in a straightforward manner to multiple benchmark groups. Let μ be fixed. We wish to show that if $\omega(y|\mu,\lambda,\sigma) = \omega(y|\mu,\tilde{\lambda},\sigma')$ for all y, then we must have $\lambda = \tilde{\lambda}$ and $\sigma = \sigma'$ (Bickel and Doksum 2015). Suppose that there exist parameters (λ,σ) and $(\tilde{\lambda},\sigma')$ such that $(\lambda,\sigma) \ne (\tilde{\lambda},\sigma')$ and $\omega(y|\mu,\lambda,\sigma) = \omega(y|\mu,\tilde{\lambda},\sigma')$ for all y. Thus, $\omega(y|\mu,\lambda,\sigma) = \omega(y|\mu,\tilde{\lambda},\sigma')$ for all $y \ge x(\bar{\theta})$ as well. Writing $\omega(y|\mu,\lambda,\sigma)$ in terms of the shock and type densities, we have $\omega(y|\mu,\lambda,\sigma) = \int_\Theta g(y - x(\theta)|\sigma,\theta)\,f(\theta|\lambda)d\theta$. Because g is the Laplace density, we have

$$\begin{aligned}&\int_\Theta (1/(2\sigma))e^{\sqrt{2}(y-x(\theta))/\sigma} f(\theta|\lambda)d\theta\\ &\quad= \int_\Theta (1/(2\sigma'))e^{\sqrt{2}(y-x(\theta))/\sigma'} f(\theta|\tilde{\lambda})d\theta \ \text{ for all } y \ge x(\bar{\theta}).\end{aligned} \tag{A.1}$$

We consider two cases: $\sigma \ne \sigma'$ and $\sigma = \sigma'$. First, suppose that $\sigma \ne \sigma'$. For conciseness, let $C = \int_\Theta (1/(2\sigma))e^{-\sqrt{2}x(\theta)/\sigma} f(\theta|\lambda)d\theta$ and $C' = \int_\Theta (1/(2\sigma'))e^{-\sqrt{2}x(\theta)/\sigma'} f(\theta|\tilde{\lambda})d\theta$, and note that C and C' are constant with respect to y. Equation (A.1) then implies that $Ce^{\sqrt{2}y/\sigma} = C'e^{\sqrt{2}y/\sigma'}$ for all $y \ge x(\bar{\theta})$. Taking the natural logarithm of both sides and rearranging yield $(\sqrt{2}/\sigma - \sqrt{2}/\sigma')y + \ln C - \ln C' = 0$ for all $y \ge x(\bar{\theta})$. This linear equation is zero over all $y \ge x(\bar{\theta})$ only if $(\sqrt{2}/\sigma - \sqrt{2}/\sigma') = 0$, which yields a contradiction. Now consider the case in which $\sigma = \sigma'$. Then we must have $\lambda \ne \tilde{\lambda}$. Cancelling out the common $(1/(2\sigma))e^{\sqrt{2}y/\sigma}$ terms on both sides of (A.1) and rearranging, we have $\int_\Theta e^{-\sqrt{2}x(\theta)/\sigma}\left(f(\theta|\lambda) - f(\theta|\tilde{\lambda})\right)d\theta = 0$, which violates the

assumption that $\int_\Theta e^{-\sqrt{2}x(\theta)/\sigma}\Big(f(\theta|\lambda) - f(\theta|\tilde{\lambda})\Big)d\theta \neq 0$ for all $\sigma \in \Sigma$ and $\lambda, \tilde{\lambda} \in \Lambda$. □

Proof of Corollary 1. Note that $\Theta = [\underline{\theta}, \bar{\theta}]$. Let $\underline{\theta}$ be fixed. Pick some $(\lambda, \tilde{\lambda}, \sigma) \in \Lambda \times \Lambda \times \Sigma$ such that $\lambda \neq \tilde{\lambda}$, and suppose that the following equality holds:

$$\int_\Theta e^{-\sqrt{2}x(\theta)/\sigma}\Big(f(\theta|\lambda) - f(\theta|\tilde{\lambda})\Big) = 0. \tag{A.2}$$

We first show that if (A.2) holds, then $\bar{\theta}$ is unique. Because $e^{-\sqrt{2}x(\theta)/\sigma}$ and $f(\theta|\lambda)$ are both strictly positive, the equality $\int_{[\theta_1,\theta_2]} e^{-\sqrt{2}x(\theta)/\sigma}\Big(f(\theta|\lambda) - f(\theta|\tilde{\lambda})\Big)d\theta = 0$ can only hold for an interval $[\theta_1, \theta_2] \subset \mathbb{R}_+$ if $f(\theta|\lambda)$ and $f(\theta|\tilde{\lambda})$ intersect at some point in (θ_1, θ_2). Further, because $f(\theta|\lambda)$ and $f(\theta|\tilde{\lambda})$ are exponential densities, they can intersect at most once over any interval (θ_1, θ_2). Now suppose that there exist multiple values of $\bar{\theta}$ such that (A.2) holds. Let $\bar{\theta}^1$ and $\bar{\theta}^2$ be two such values, where $\bar{\theta}^1 < \bar{\theta}^2$. Then we have $\int_{[\underline{\theta},\bar{\theta}^1]} e^{-\sqrt{2}x(\theta)/\sigma}\Big(f(\theta|\lambda) - f(\theta|\tilde{\lambda})\Big)d\theta = 0$ and $\int_{[\underline{\theta},\bar{\theta}^2]} e^{-\sqrt{2}x(\theta)/\sigma}\Big(f(\theta|\lambda) - f(\theta|\tilde{\lambda})\Big)d\theta = 0$, which imply that $\int_{[\bar{\theta}^1,\bar{\theta}^2]} e^{-\sqrt{2}x(\theta)/\sigma}\Big(f(\theta|\lambda) - f(\theta|\tilde{\lambda})\Big)d\theta = 0$ as well. This implies that $f(\theta|\lambda)$ and $f(\theta|\tilde{\lambda})$ intersect in both $(\underline{\theta}, \bar{\theta}^1)$ and $(\bar{\theta}^1, \bar{\theta}^2)$, a contradiction. Thus, there is at most one value $\bar{\theta}$ such that (A.2) holds. Let this parameter be $\bar{\theta}(\lambda, \tilde{\lambda}, \sigma)$. Because Λ and Σ are discrete, there are a finite number of such $\bar{\theta}(\lambda, \tilde{\lambda}, \sigma)$. Selecting $\bar{\theta} > \sup_{\lambda,\tilde{\lambda}\in\Lambda,\sigma\in\Sigma} \bar{\theta}(\lambda, \tilde{\lambda}, \sigma)$ implies that (A.2) cannot hold for any $(\lambda, \tilde{\lambda}, \sigma) \in \Lambda \times \Lambda \times \Sigma$, which yields the result. □

Proof of Proposition 5. For conciseness, let $\lambda = (\lambda_1, \ldots, \lambda_m)$ in what follows. Let $\omega(y|\mu, \lambda, \sigma)$ be the savings density for an ACO with benchmark μ, given λ and σ. Letting $\mathscr{L}(\lambda, \sigma|\mu, y)$ be the likelihood function, we can write

$$\begin{aligned}
\mathscr{L}(\lambda, \sigma|\mu, y) &= \prod_{i=1}^n \omega(y_i|\mu_i, \lambda, \sigma) \\
&= \prod_{i=1}^n \int_\Theta \omega(y_i|\mu_i, \lambda, \sigma, \theta) f(\theta|\mu_i, \lambda, \sigma) d\theta \\
&= \prod_{i=1}^n \int_\Theta g(y_i - x(\theta)|\mu_i, \lambda, \sigma, \theta) f(\theta|\mu_i, \lambda, \sigma) d\theta \\
&= \prod_{i=1}^n \int_\Theta g(y_i - x(\theta)|\mu_i, \lambda(\mu_i), \sigma, \theta) f(\theta|\lambda(\mu_i)) d\theta \\
&= \prod_{i=1}^n \int_\Theta g(y_i - x(\theta)|\sigma, \theta) f(\theta|\lambda(\mu_i)) d\theta.
\end{aligned}$$

The first line follows by definition of the likelihood function and the independence assumption given in Assumption 5. The second line follows from conditioning on θ. The third line follows from noting that $y = x(\theta) + \xi$ and rewriting the savings distribution in terms of the shock density g. The fourth line follows from noting that $f(\theta|\mu_i, \lambda, \sigma) = f(\theta|\lambda(\mu_i))$ because $\lambda(\mu_i)$ fully defines the type density of an ACO. The final line follows by observing that $y_i - x(\theta)$ depends only on σ and θ (in addition to α and β, which are fixed throughout). Taking the logarithm of both sides, we obtain

$$\log \mathscr{L}(\lambda, \sigma|\mu, y) = \sum_{i=1}^n \log\left(\int_\Theta g(y_i - x(\theta)|\sigma, t) f(t|\lambda(\mu_i)) d\theta\right),$$

as desired. □

Endnotes

[1] The MSSP is distinct (but not mutually exclusive) from the Bundled Payments for Care Improvement (BPCI) Initiative, another recently established Medicare program that incentivizes providers to reduce the cost of healthcare delivery (CMS 2017b). The BPCI Initiative offers Medicare providers a single reimbursement for "bundles" of healthcare services received by a beneficiary during a single episode of care, in lieu of reimbursing the provider for each individual service provided. In contrast to the MSSP, participation in the BPCI Initiative does not require the formation of an ACO or include financial bonuses in the form of shared savings payments. Unlike the MSSP, the BPCI Initiative is relatively well studied in the operations management literature (Gupta and Mehrotra 2015, Adida et al. 2016, Guo et al. 2016).

[2] To keep the the model tractable and amenable to estimation, we assume that the provider's service-related profit decreases with its cost reduction efforts, but the profit margin itself remains constant at γ. This is supported by the notion that the MSSP aims, in part, to generate savings by reducing the total number of healthcare services provided, meaning the average per-service profit margin will not necessarily be impacted by the provider's operational decisions.

[3] In general, we assume that the ACO participates if its payoff is nonnegative. In our analysis, however, we shall use a stronger constraint to guarantee that the ACO's payoff is no less than it would be under the existing MSSP contract (i.e., strictly positive). We impose this stronger condition to restrict attention to contracts that improve Medicare savings without leaving the ACO worse off.

[4] A limitation of our work is that our formulation of Medicare's contracting problem takes a single-period view of the MSSP. In a multiperiod setting, however, formulation OC-I implies that asymmetric information regarding the ACO's type may not persist after the first period because the ACO is incentivized to immediately and truthfully report its type. If the ACO's type does not change, then the revelation of the ACO's type in the first period substantively changes the nature of Medicare's contracting problem in subsequent periods, which we do not address in this paper.

[5] In practice, the parameters C_u and $|C_\ell|$ are large relative to typical ACO savings. According to existing MSSP guidelines, C_ℓ and C_u are set at 15% and 20% of the ACO's benchmark, respectively (Federal Register 2011), whereas data released by the CMS show that the majority of ACO savings and losses are within 5% of the benchmark.

[6] The "optimal nonsubsidized contract" refers to the existing MSSP structure under the optimal shared savings parameters $\alpha(\theta)$, where $\beta(\theta) = 0$ for all $\theta \in \Theta$.

[7] Because we have only a single observation for each ACO, we cannot estimate the type parameter θ for each individual ACO in the data set. We therefore focus on estimating the distribution over ACO types in aggregate. However, it may be possible to estimate the exact type parameter for each ACO given multiple observations of the same ACO over several years. This may be a potentially fruitful direction for future analyses of the MSSP.

[8] We estimate the type distribution $f(\theta)$ by using the normalized per-beneficiary savings of each ACO. In other words, to reduce model complexity, we do not explicitly account for variation in ACO size in the estimator. We instead capture variation in ACO size when simulating contract performance by sampling the number of beneficiaries along with other ACO attributes in the bootstrap.

[9] In Section EC.3 of the online supplemental material, we consider an extension whereby the threshold h is also optimized in addition to the shared savings rate α and the subsidy rate β.

[10] Note that using a finer discretization for the contract and type space may improve the performance of the contract by more tightly approximating the original optimal contracting problem. However, because we formulate the optimal contracting problems as integer

optimization models, the models become intractably large if the discretization scheme is too fine. Note also that because of the discretization, our estimates of Medicare's savings under the optimal contract are likely to be conservative. The development of efficient solution techniques for optimal contracting problems that are formulated as large-scale integer optimization models may be a fruitful direction for future work.

References

Adida E, Mamani H, Nassiri S (2016) Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Sci.* 63(5):1606–1624.

Ahuja RK, Orlin JB (2001) Inverse optimization. *Oper. Res.* 49(5): 771–783.

Akira Health, Inc. (2017) Akira health accountable care organization. Accessed June 1, 2018, http://www.akirahealth.com/.

Anderson GF, Steinberg EP (1984) Hospital readmissions in the Medicare population. *New England J. Medicine* 311(21):1349–1353.

Andritsos D, Tang CS (2015) Incentive programs for reducing readmissions when patient care is co-produced. *Production Oper. Management* 27(6):999–1020.

Arifoglu K, Deo S, Iravani SMR (2012) Consumption externality and yield uncertainty in the influenza vaccine supply chain: Interventions in demand and supply sides. *Management Sci.* 58(6): 1072–1091.

Aswani A, Max Shen Z-J, Siddiq A (2018) Inverse optimization with noisy data. *Oper. Res.* 66(3):870–892.

Ata B, Killaly BLParker RP, Olsen TL (2013) On hospice operations under medicare reimbursement policies. *Management Sci.* 59(5): 1027–1044.

Ata B, Lee D, Tongarlak MH (2012) Optimizing organic waste to energy operations. *Manufacturing Service Oper. Management* 14(2): 231–244.

Balasubramanian S, Bhardwaj P (2004) When not all conflict is bad: Manufacturing-marketing conflict and strategic incentive design. *Management Sci.* 50(4):489–502.

Bastani H, Bayati M, Braverman M, Gummadi R, Johari R (2016) Analysis of medicare pay-for-performance contracts. Working paper, University of Pennsylvania, Philadelphia.

Bertsimas D, Gupta V, Ch Paschalidis I (2015) *Data-Driven Estimation in Equilibrium Using Inverse Optimization*. *Mathematical Programming Series A*, vol. 153 (Springer-Verlag, Berlin, Heidelberg), 595–633.

Berwick DM (2011) Launching accountable care organizations: The proposed rule for the medicare shared savings program. *New England J. Medicine* 364(16):e32.

Berwick DM, Hackbarth AD (2012) Eliminating waste in US health care. *J. Amer. Medical Assoc.* 307(14):1513–1516.

Bickel PJ, Doksum KA (2015) *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. 2 (CRC Press, Boca Raton, FL).

Borgers T, Strausz R, Krahmer D (2015) *An Introduction to the Theory of Mechanism Design* (Oxford University Press, New York).

Chemama J, Cohen MC, Lobel R, Perakis G (2019) Consumer subsidies with a strategic supplier: Commitment vs. flexibility. *Management Sci.* 65(2):681–713.

Chen F (2000) Sales-force incentives and inventory management. *Manufacturing Service Oper. Management* 2(2):186–202.

Chen S, Lee H (2016) Incentive alignment and coordination of project supply chains. *Management Sci.* 63(4):1011–1025.

Chen T, Klastorin T, Wagner MR (2015) Incentive contracts in serial stochastic projects. *Manufacturing Service Oper. Management* 17(3) 290–301.

Chernew M, McGuire T, McWilliams JM (2014) *Refining the ACO Program: Issues and Options* (Department of Healthcare Policy, Harvard Medical School, Boston).

Chick SE, Mamani H, Simchi-Levi D (2008) Supply chain coordination and influenza vaccination. *Oper. Res.* 56(6):1493–1506.

CMS (2016a) 2016 Press release: Physicians and health care providers continue to improve quality of care, lower costs. Accessed September 1, 2016, https://www.cms.gov/newsroom/press-releases/physicians-and-health-care-providers-continue-improve-quality-care-lower-costs.

CMS (2016b) About the program: Centers for Medicare and Medicaid Services. Accessed June 1, 2016, https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/about.html.

CMS (2016c) ACO investment model: Centers for Medicare and Medicaid Services. Accessed June 1, 2016, https://innovation.cms.gov/initiatives/ACO-_Investment-_Model/.

CMS (2016d) Program guidelines and specifications: Improving quality of care for medicare patients: Accountable care organizations. Accessed August 1, 2018, https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/program-guidance-and-specifications.html.

CMS (2017a) Medicare shared savings program accountable care organizations performance year 2015 results. Accessed August 1, 2018, https://data.cms.gov/Special-Programs-Initiatives-Medicare-Shared-Savin/Medicare-Shared-Savings-Program-Accountable-Care-O/x8va-z7cu.

CMS (2017b) Bundled payments for care improvement (BPCI) initiative: General information. Accessed August 1, 2018, https://innovation.cms.gov/initiatives/bundled-_payments/.

Cohen MC, Lobel R, Perakis G (2015a) The impact of demand uncertainty on consumer subsidies for green technology adoption. *Management Sci.* 62(5):1235–1258.

Cohen MC, Perakis G, Thraves C (2015b) Competition and externalities in green technology adoption. Working paper, New York University, New York.

Crosson FJ (2011) The accountable care organization: Whatever its growing pains, the concept is too vitally important to fail. *Health Affairs* 30(7):1250–1255.

DeHoratius N, Raman A (2007) Store manager incentive design and retail performance: An exploratory investigation. *Manufacturing Service Oper. Management* 9(4):518–534.

Dempe S (2002) *Foundations of Bilevel Programming* (Springer Science & Business Media, New York).

Douven R, McGuire TG, McWilliams JM (2015) Avoiding unintended incentives in ACO payment models. *Health Affairs (Millwood)* 34(1):143–149.

Eddy DM, Shah R (2012) A simulation shows limited savings from meeting quality targets under the Medicare shared savings program. *Health Affairs (Millwood)* 31(11):2554–2562.

Efron B, Tibshirani RJ (1994) *An Introduction to the Bootstrap* (CRC Press, Boca Raton, FL).

Federal Register (2011) Medicare shared savings program: Accountable care organizations. Final rule. *Federal Register* 76(212):67802.

Fisher ES, Shortell SM, Kreindler SA, Van Citters AD, Larson BK (2012) A framework for evaluating the formation, implementation, and performance of accountable care organizations. *Health Affairs (Millwood)* 31(11):2368–2378.

Fuloria PC, Zenios SA (2001) Outcomes-adjusted reimbursement in a health-care delivery system. *Management Sci.* 47(6):735–751.

Garber AM, Skinner J (2008) Is American health care uniquely inefficient? Technical report, National Bureau of Economic Research, Cambridge, MA.

Gibbons R (1998) Incentives in organizations. Technical report, National Bureau of Economic Research, Cambridge, MA.

Grinblatt M, Titman S (1989) Adverse risk incentives and the design of performance-based contracts. *Management Sci.* 35(7):807–822.

Grossman SJ, Hart OD (1983) An analysis of the principal-agent problem. *Econometrica* 51(1):7–45.

Guajardo JA, Cohen MA, Kim S-H, Netessine S (2012) Impact of performance-based contracting on product reliability: An empirical analysis. *Management Sci.* 58(5):961–979.

Guo P, Tang CS, Wang Y, Zhao M (2016) The impact of reimbursement policy on patient welfare, readmission rate and waiting time in a public healthcare system: Fee-for-service vs bundled payment. Working paper, London Business School, London.

Gupta D, Mehrotra M (2015) Bundled payments for healthcare services: Proposer selection and information sharing. *Oper. Res.* 63(4):772–788.

Harris M, Raviv A (1979) Optimal incentive contracts with imperfect information. *J. Econom. Theory* 20(2):231–259.

Hart OD, Holmström B (1986) *The Theory of Contracts* (Department of Economics, Massachusetts Institute of Technology, Cambridge, MA).

Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: Data mining, inference and prediction. *Math. Intelligencer* 27(2):83–85.

Haywood TT, Kosel KC (2011) The ACO model—A three-year financial loss? *New England J. Medicine* 364(14):e27.

Hendee WR, Becker GJ, Borgstede JP, Bosma J, Casarella WJ, Erickson BA, Maynard CD, Thrall JH, Wallner PE (2010) Addressing overutilization in medical imaging. *Radiology* 257(1):240–245.

Hölmstrom B (1979) Moral hazard and observability. *Bell J. Econom.* 10(1):74–91.

Jiang H, Pang S, Savin S (2016) Capacity management for outpatient medical services under competition and performance-based incentives. Working paper, Wharton School, University of Pennsylvania, Philadelphia.

Jiang H, Zhan P, Savin S (2012) Performance-based contracts for outpatient medical services. *Manufacturing Service Oper. Management* 14(4):654–669.

Khanjari NE, Iravani S, Shin H (2013) The impact of the manufacturer-hired sales agent on a supply chain with information asymmetry. *Manufacturing Service Oper. Management* 16(1):76–88.

Khouja M, Zhou J (2010) The effect of delayed incentives on supply chain profits and consumer surplus. *Production Oper. Management* 19(2):172–197.

Laffont J-J, Martimort D (2009) *The Theory of Incentives: The Principal-Agent Model* (Princeton University Press, Princeton, NJ).

Lariviere MA (2015) OM forum—Supply chain contracting: doughnuts to bubbles. *Manufacturing Service Oper. Management* 18(3):309–313.

Lee DKK, Zenios SA (2012) An evidence-based incentive system for Medicare's end-stage renal disease program. *Management Sci.* 58(6):1092–1105.

Levi R, Perakis G, Romero G (2016) On the effectiveness of uniform subsidies in increasing market consumption. *Management Sci.* 63(1):40–57.

Lieberman SM, Bertko JM (2011) Building regulatory and operational flexibility into accountable care organizations and 'shared savings.' *Health Affairs (Millwood)* 30(1):23–31.

Liu P, Wu S (2014) An agent-based simulation model to study accountable care organizations. *Health Care Management Sci.* 19(1):89–101.

Mamani H, Chick SE, Simchi-Levi D (2013) A game-theoretic model of international influenza vaccination coordination. *Management Sci.* 59(7):1650–1670.

Massey FJ (1951) The Kolmogorov-smirnov test for goodness of fit. *J. Amer. Statist. Assoc.* 46(253):68–78.

McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, Kerr EA (2003) The quality of health care delivered to adults in the United States. *New England J. Medicine* 348(26):2635–2645.

McWilliams JM, Chernew ME, Landon BE, Schwartz AL (2015) Performance differences in year 1 of pioneer accountable care organizations. *New England J. Medicine* 372(20):1927–1936.

McWilliams JM, Hatfield LA, Chernew ME, Landon BE, Schwartz AL (2016) Early performance of accountable care organizations in medicare. *New England J. Medicine* 374(24):2357–2366.

McWilliams JM, Landon BE, Chernew ME (2013) Changes in health care spending and quality for Medicare beneficiaries associated with a commercial ACO contract. *J. Amer. Medical Assoc.* 310(8):829–836.

MedPAC (2010) Healthcare spending and the medicare program. Medicare Payment Advisory Commission. Accessed August 1, 2018, http://67.59.137.244/documents/Jun10DataBookEntireReport.pdf.

Milgate K, Cheng SB (2006) Pay-for-performance: The MedPAC perspective. *Health Affairs (Millwood)* 25(2):413–419.

Miller HD (2009) From volume to value: Better ways to pay for health care. *Health Affairs (Millwood)* 28(5):1418–1428.

Moore KD (2011) *The Work Ahead: Activities and Costs to Develop an Accountable Care Organization* (American Hospital Association, Chicago).

Myerson RB (1981) Optimal auction design. *Math. Oper. Res.* 6(1):58–73.

National Association of ACOs (2014) National ACO survey. Accessed June 1, 2016, https://www.naacos.com/assets/docs/pdf/acosurveyfinal012114.pdf.

OECD (2016) Organization for economic cooperation and development health statistics. Accessed September 1, 2018, http://www.oecd.org/els/health-systems/health-data.htm.

Plambeck EL, Zenios SA (2000) Performance-based incentives in a dynamic principal-agent model. *Manufacturing Service Oper. Management* 2(3):240–263.

Raghu TS, Sen PK, Rao HR (2003) Relative performance of incentive mechanisms: Computational modeling and simulation of delegated investment decisions. *Management Sci.* 49(2):160–178.

Rosenthal MB, Cutler DM, Feder J (2011) The ACO rules—Striking the balance between participation and transformative potential. *New England J. Medicine* 365(4):e6.

Rosenthal MB, Fernandopulle R, Ryu Song HS, Landon B (2004) Paying for quality: Providers' incentives for quality improvement. *Health Affairs (Millwood)* 23(2):127–141.

Savva N, Tezcan T, Yildiz O (2016) Yardstick competition for service systems. Working paper, London Business School, London.

Shavell S (1979a) On moral hazard and insurance. Dionne G, Harrington SE, eds. *Foundations of Insurance Economics* (Springer, New York), 280–301.

Shavell S (1979b) Risk sharing and incentives in the principal and agent relationship. *Bell J. Econom.* 10(1):55–73.

So KC, Tang CS (2000) Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Sci.* 46(7):875–892.

Starfield B (2000) Is US health really the best in the world? *J. Amer. Medical Assoc.* 284(4):483–485.

Taylor TA, Xiao W (2014) Subsidizing the distribution channel: Donor funding to improve the availability of malaria drugs. *Management Sci.* 60(10):2461–2477.

Wennberg JE, Fisher ES, Skinner JS (2002) Geography and the debate over Medicare reform. *Health Affairs (Millwood)* 21(2):10.

Whang S (1992) Contracting for software development. *Management Sci.* 38(3):307–324.

Wilensky GR (2013) Developing a viable alternative to Medicare's physician payment strategy. *Health Affairs (Millwood)* 33(1):153–160.

Yamin D, Gavious A (2013) Incentives' effect in influenza vaccination policy. *Management Sci.* 59(12):2667–2686.

Zhang D, Gurvich I, Van Mieghem J, Park E, Young R, Williams M (2016a) Hospital readmissions reduction program: An economic and operational analysis. *Management Sci.* 62(11):3351–3371.

Zhang H, Wernz C, Hughes DR (2016b) Modeling and designing health care payment innovations for medical imaging. *Health Care Management Sci.* 21(1):37–51.

Anil Aswani is an assistant professor in the Department of Industrial Engineering and Operations Research at the University of California, Berkeley. His research interests include data-driven decision making, with particular emphasis on

addressing inefficiencies and inequities in health systems and physical infrastructure.

Zuo-Jun (Max) Shen is a chancellor's professor in the Department of Industrial Engineering and Operations Research and the Department of Civil and Environmental Engineering at University of California, Berkeley. He is also an honorary professor at Tsinghua University. He has been active in the following research areas: integrated supply-chain design and management, design and analysis of optimization algorithms, energy system and transportation system planning, and optimization.

Auyon Siddiq is an assistant professor in the Anderson School of Management at the University of California, Los Angeles. His research interests include data analytics, operations management, and policy issues.

Electronic companion for "Data-Driven Incentive Design in the Medicare Shared Savings Program".

EC.1. Validation of Assumptions 3 and 4

Here we show that Assumptions 3 and 4 are validated by the dataset. First we provide values of the model parameters based on the data such that Assumption 3 holds. Specifically, we wish to show that $\bar{\alpha}(h\bar{g}' + \bar{g}) + 2\bar{\beta}\bar{g}(\partial/\partial x)c(x,\theta) \leq (1-\bar{\beta})(\partial^2/\partial x^2)c(x,\theta)$ holds for all $x \in [0,\bar{x}]$ and $\theta \in \Theta$. We shall consider the model and parameter estimates from Sections 6 and 7. Since we use $c(x,\theta) = x^2/\theta$ for the ACO investment function, we have $(\partial/\partial x)c(x,\theta) = 2x/\theta$ and $(\partial^2/\partial x^2)c(x,\theta) = 2/\theta$. Let $\bar{\alpha} = 0.55$, $\bar{\beta} = 0.05$, $\bar{x} = 1{,}000$, $\underline{\theta} = 1$ and $\bar{\theta} = 2{,}000$. Based on the guidelines of the MSSP, the parameter h is approximately 200 (2% of the benchmark). Since the shock density g is assumed to be Laplace distributed with an estimated standard deviation of $\hat{\sigma} = 550$, this corresponds to $\bar{g} = 1/(\sqrt{2}\hat{\sigma}) = 0.0013$ and $\bar{g}' = (1/\hat{\sigma}^2) = 2.8 \times 10^{-6}$. Using these parameter values, we can numerically verify that the inequality holds for all $x \in [0,\bar{x}]$ and $\theta \in \Theta$. Next, for Assumption 4, we have

$$\begin{aligned}(\partial/\partial x)R(x,\alpha) &= (1-\alpha)\left(hg(-h-x) + G(-h-x) - G(C_\ell - x)\right) + \alpha\left(hg(h-x) + G(C_u - x) - G(h-x)\right) \\ &< (1-\alpha)\left(hg(-h-x) + 1/2\right) + \alpha\left(hg(h-x) + 1\right) \\ &\leq (1-\alpha)\left(hg(-h) + 1/2\right) + \alpha\left(h\bar{g} + 1\right)\end{aligned}$$

The first line is given by the expression for $(\partial/\partial x)R(x,\alpha)$ given in Lemma 4. The second line follows from noting that $G(-h-x) - G(C_\ell - x) \leq 1/2$ (since $-h-x < 0$) and $G(C_u - x) - G(h-x) \leq 1$. The third line follows from the fact that $\bar{g} \geq g(h-x)$ and $g(-h) \geq g(-h-x)$ for $x \geq 0$. For Assumption 4 to hold, we require $(1-\alpha)\left(hg(-h) + 1/2\right) + \alpha\left(h\bar{g} + 1\right) < 1$ for all $\alpha \in \mathcal{A}$, which holds if $\bar{\alpha} \leq (1 - 2hg(-h))/(1 + 2h(\bar{g} - g(-h)))$. The inequality can be shown to hold for $\bar{\alpha} = 0.55$, $h = 200$ and $\hat{\sigma} = 550$, where $\bar{g} = 1/(\sqrt{2}\hat{\sigma})$ and $g(-h) = 1/(\sqrt{2}\hat{\sigma})e^{\sqrt{2}|-h|/\hat{\sigma}}$.

EC.2. Integer Optimization Models for Medicare's Optimal Contracting Problem

This section presents four integer optimization models for the optimal contracting problems OC-I, OC-II, OC-III, and OC-IV, which are given in Sections 4 and 5. First, we consider OC-I and OC-II. Let $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_{|\mathcal{I}|}\}$ be the set of shared savings rates, $\mathcal{B} = \{\beta_1, \beta_2, \ldots, \beta_{|\mathcal{J}|}\}$ be the set of subsidy-rates, and $\Theta = \{\theta_1, \theta_2, \ldots, \theta_{|\mathcal{K}|}\}$ be the set of ACO types. Let $u(\alpha_i, \beta_j, \theta_k)$ be the optimal payoff of a type θ_k ACO under contract parameters α_i and β_j. Note that given $(\alpha_i, \beta_j, \theta_k)$, $u(\alpha_i, \beta_j, \theta_k)$ can be computed by solving the ACO's investment problem (represented by (4) and (5)) using simple line search techniques. Similarly, let $v(\alpha_i, \beta_j, \theta_k)$ be Medicare's savings under parameters $(\alpha_i, \beta_j, \theta_k)$. As in formulation OC-I, let $\bar{u}(\theta_k)$ be the minimum pay-off for a type θ_k ACO. Let $p(\theta_k)$

be the probability that an ACO is type θ_k, where $p(\theta_k)$ can be computed by discretizing the type distribution estimated in Section 6. The key decision variable in the integer optimization model is the binary variable $z(\alpha_i, \beta_j, \theta_k)$, where $z(\alpha_i, \beta_j, \theta_k) = 1$ if the contract parameters α_i and β_j are mapped to a type θ_k ACO, and 0 otherwise. The optimal contracting problem OC-I can then be formulated as the following integer optimization model:

$$\underset{\mathbf{z}}{\text{maximize}} \quad \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} v(\alpha_i,\beta_j,\theta_k)\cdot z(\alpha_i,\beta_j,\theta_k)\cdot p(\theta_k) \tag{EC.1a}$$

$$\text{subject to} \quad \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} u(\alpha_i,\beta_j,\theta_k)\cdot z(\alpha_i,\beta_j,\theta_k) \geq \bar{u}(\theta_k), \quad k\in\mathcal{K}, \tag{EC.1b}$$

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} u(\alpha_i,\beta_j,\theta_k)\cdot z(\alpha_i,\beta_j,\theta_k) \geq \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} u(\alpha_i,\beta_j,\theta_k)\cdot z(\alpha_i,\beta_j,\theta_{k'}), \quad k,k'\in\mathcal{K}, \tag{EC.1c}$$

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} z(\alpha_i,\beta_j,\theta_k) = 1, \quad k\in\mathcal{K}, \tag{EC.1d}$$

$$z(\alpha_i,\beta_j,\theta_k)\in\{0,1\}, \quad i\in\mathcal{I},\ j\in\mathcal{J},\ k\in\mathcal{K}. \tag{EC.1e}$$

Observe that (EC.1a), (EC.1b) and (EC.1c) are the discrete counterparts to (7a), (7b) and (7c). Constraint (EC.1d) ensures that exactly one set of contract parameters is assigned to each ACO type. To formulate OC-II as an integer optimization problem, we require additional integer decision variable to represent the indicator variable in the objective function of OC-II. Let $\zeta_1(\alpha_i,\beta_j,\theta_k)$ and $\zeta_2(\alpha_i,\beta_j,\theta_k)$ be binary variables. The integer optimization model for problem OC-II is given by

$$\underset{\mathbf{z},\zeta_1,\zeta_2}{\text{maximize}} \quad \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} v(\alpha_i,\beta_j,\theta_k)\cdot \zeta^1(\alpha_i,\beta_j,\theta_k)\cdot p(\theta_k) \tag{EC.2a}$$

$$\text{subject to} \quad \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} u(\alpha_i,\beta_j,\theta_k)\cdot z(\alpha_i,\beta_j,\theta_k) \geq \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} u(\alpha_i,\beta_j,\theta_k)\cdot z(\alpha_i,\beta_j,\theta_{k'}), \quad k,k'\in\mathcal{K}, \tag{EC.2b}$$

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} z(\alpha_i,\beta_j,\theta_k) = 1, \quad k\in\mathcal{K}, \tag{EC.2c}$$

$$u(\alpha_i,\beta_j,\theta_k) \leq C_u\cdot\zeta_2(\alpha_i,\beta_j,\theta_k), \quad i\in\mathcal{I}, j\in\mathcal{J}, k\in\mathcal{K}, \tag{EC.2d}$$

$$-u(\alpha_i,\beta_j,\theta_k) \leq C_u\cdot(1-\zeta_2(\alpha_i,\beta_j,\theta_k)), \quad i\in\mathcal{I}, j\in\mathcal{J}, k\in\mathcal{K}, \tag{EC.2e}$$

$$2\zeta_1(\alpha_i,\beta_j,\theta_k) \leq \zeta_2(\alpha_i,\beta_j,\theta_k) + z(\alpha_i,\beta_j,\theta_k), \quad i\in\mathcal{I}, j\in\mathcal{J}, k\in\mathcal{K}, \tag{EC.2f}$$

$$\zeta_2(\alpha_i,\beta_j,\theta_k) + z(\alpha_i,\beta_j,\theta_k) \leq 1 + \zeta_1(\alpha_i,\beta_j,\theta_k), \quad i\in\mathcal{I}, j\in\mathcal{J}, k\in\mathcal{K}, \tag{EC.2g}$$

$$z(\alpha_i,\beta_j,\theta_k),\ \zeta_1(\alpha_i,\beta_j,\theta_k),\ \zeta_2(\alpha_i,\beta_j,\theta_k)\in\{0,1\}, \quad i\in\mathcal{I},\ j\in\mathcal{J},\ k\in\mathcal{K}. \tag{EC.2h}$$

Constraints (EC.2d) and (EC.2e) force $\zeta^2(\alpha_i,\beta_j,\theta_k) = 1$ if and only if $u(\alpha_i,\beta_j,\theta_k) > 0$. Constraints (EC.2f) and (EC.2g) force $\zeta^1(\alpha_i,\beta_j,\theta_k) = 1$ if and only if $\zeta^2(\alpha_i,\beta_j,\theta_k) = 1$ and $z(\alpha_i,\beta_j,\theta_k) = 1$. Therefore, $\zeta^1(\alpha_i,\beta_j,\theta_k) = 1$ if the contract parameters α_i and β_j are mapped to a type θ_k ACO *and* if $u(\alpha_i,\beta_j,\theta_k) \geq 0$, and 0 otherwise. The integer optimization model for the non-parametric contracts

can be formulated in a similar manner. Let $\{\rho_1, \rho_2, \ldots, \rho_{|\mathcal{L}|}\}$,$\{x_1, x_2, \ldots, x_{|\mathcal{M}|}\}$, and $\{\theta_1, \theta_2, \ldots, \theta_{|\mathcal{K}|}\}$ be the set of possible payments, ACO savings levels, and ACO types, respectively. By the revelation principle (Myerson 1981), it suffices to restrict attention to incentive-compatible contracts that map a payment ρ and savings x to each ACO type θ. The decision variable in the non-parametric formulation is thus $w(\rho_l, x_m, \theta_k)$, where $w(\rho_l, x_m, \theta_k) = 1$ if the payment–savings pair (ρ_l, x_m) is assigned to a type θ_k ACO, and 0 otherwise. The integer optimization model for contracting problem OC-III is then given by

$$\underset{\mathbf{w}}{\text{maximize}} \quad \sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} v(\rho_l, x_m, \theta_k) \cdot w(\rho_l, x_m, \theta_k) \cdot p(\theta_k) \tag{EC.3a}$$

$$\text{subject to} \quad \sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} u(\rho_l, x_m, \theta_k) \cdot z(\rho_l, x_m, \theta_k) \geq \bar{u}(\theta_k), \quad k \in \mathcal{K}, \tag{EC.3b}$$

$$\sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} u(\rho_l, x_m, \theta_k) \cdot w(\rho_l, x_m, \theta_k) \geq \sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} u(\rho_l, x_m, \theta_k) \cdot z(\rho_l, x_m, \theta_{k'}), \quad k, k' \in \mathcal{K}, \tag{EC.3c}$$

$$\sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} w(\rho_l, x_m, \theta_k) = 1, \quad k \in \mathcal{K}, \tag{EC.3d}$$

$$w(\rho_l, x_m, \theta_k) \in \{0, 1\}, \quad l \in \mathcal{L},\ m \in \mathcal{M},\ k \in \mathcal{K}. \tag{EC.3e}$$

Lastly, the integer optimization model for problem OC-IV is

$$\underset{\mathbf{w},\eta_1,\eta_2}{\text{maximize}} \quad \sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} \sum_{k\in\mathcal{K}} v(\rho_l, x_m, \theta_k) \cdot \eta_1(\rho_l, x_m, \theta_k) \cdot p(\theta_k) \tag{EC.4a}$$

$$\text{subject to} \quad \sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} u(\rho_l, x_m, \theta_k) \cdot w(\rho_l, x_m, \theta_k) \geq \sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} u(\rho_l, x_m, \theta_k) \cdot z(\rho_l, x_m, \theta_{k'}), \quad k, k' \in \mathcal{K}, \tag{EC.4b}$$

$$\sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} w(\rho_l, x_m, \theta_k) = 1, \quad k \in \mathcal{K}, \tag{EC.4c}$$

$$u(\rho_l, x_m, \theta_k) \leq \rho_{|\mathcal{L}|} \cdot \eta_2(\rho_l, x_m, \theta_k), \quad l \in \mathcal{L}, m \in \mathcal{M}, k \in \mathcal{K}, \tag{EC.4d}$$

$$-u(\rho_l, x_m, \theta_k) \leq \rho_{|\mathcal{L}|} \cdot (1 - \eta_2(\rho_l, x_m, \theta_k)), \quad l \in \mathcal{L}, m \in \mathcal{M}, k \in \mathcal{K}, \tag{EC.4e}$$

$$2\eta_1(\rho_l, x_m, \theta_k) \leq \eta_2(\rho_l, x_m, \theta_k) + z(\rho_l, x_m, \theta_k), \quad l \in \mathcal{L}, m \in \mathcal{M}, k \in \mathcal{K}, \tag{EC.4f}$$

$$\eta_2(\rho_l, x_m, \theta_k) + z(\rho_l, x_m, \theta_k) \leq 1 + \eta_1(\rho_l, x_m, \theta_k), \quad l \in \mathcal{L}, m \in \mathcal{M}, k \in \mathcal{K}, \tag{EC.4g}$$

$$w(\rho_l, x_m, \theta_k), \eta_1(\rho_l, x_m, \theta_k), \eta_2(\rho_l, x_m, \theta_k) \in \{0, 1\}, \quad l \in \mathcal{L},\ m \in \mathcal{M},\ k \in \mathcal{K}. \tag{EC.4h}$$

In this model, the variables $\eta_1(\alpha_i, \beta_j, \theta_k)$ and $\eta_2(\alpha_i, \beta_j, \theta_k)$ assume parallel meanings to $\zeta_1(\alpha_i, \beta_j, \theta_k)$ and $\zeta_2(\alpha_i, \beta_j, \theta_k)$ from formulation (EC.2). All four of the models above are pure binary integer programs, and can therefore be solved using off-the-shelf optimization solvers. The optimal subsidized and non-parametric contracts can be recovered from the optimal assignment vectors $\mathbf{z}^*$ and $\mathbf{w}^*$, respectively.

EC.3. Extensions and Sensitivity Tests

In this section we consider four extensions to our main model (OC-I). First, we consider Medicare's optimal contracting problem when the minimum savings threshold h is optimized in addition to the shared savings rate α and the subsidy rate β. Second, we repeat the empirical analysis (parameter estimation, contract optimization and bootstrap simulation) under two alternate specifications for the ACO's investment function $c(x,\theta)$, and compare these results with our findings from Section 7. Third, we numerically analyze a setting where there is inaccuracy in the financial benchmark, i.e., Medicare observes a noisy signal of μ instead of μ directly. Fourth, we consider an alternate contract that only depends on the ACO's benchmark μ, instead of θ.

Optimizing savings threshold. We first consider a setting where the threshold parameter h is also a decision variable in addition to α and β. Under the current MSSP contract structure, $h = \mathsf{MSR} \times \mu$, where μ is the ACO's benchmark and MSR is the *minimum savings rate*, expressed as a percentage. Under the existing contract, $\mathsf{MSR} = 2\%$ Federal Register (2011). To maintain tractability of the optimal contracting problem, we vary MSR from 0% to 4% in increments of 0.25%, where for each value of MSR, we re-solve the optimal contracting problem OC-I (via its discrete counterpart given in formulation (EC.1)) and repeat the bootstrap procedure to estimate Medicare's savings under the associated optimal contract. The estimated savings for varying values of MSR are shown in Table EC.1. The optimal threshold based on the simulation is 1.75%, which leads to a slight improvement in Medicare savings compared to the default value of 2.00% ($210 vs $207 million).

Alternate investment functions. Table EC.2 presents the estimated savings under the baseline and optimal contracts for three different investment functions, including $c(x,\theta) = x^2/\theta$, which is used in our main analysis in Sections 6. Under the current two-sided contract, we estimate the total Medicare savings to range from $135 to $146 million. Our estimates of Medicare's savings under the optimal subsidy contract range from $187 to $383 million. Given that total spending by this group of ACOs was approximately $73 billion, these results suggest that our estimates of Medicare savings under the optimal contract are reasonably robust to the choice of the investment function. To measure the goodness-of-fit for each of the three specifications of the investment function, we compute the Kolmogorov-Smirnov statistic (Massey 1951) for the empirical and simulated savings distributions under each of the three models (where a lower test statistic suggests a better fit). The test statistics for x^2/θ, $(x^2+x)/\theta$, and $(x/\theta)\log(x+1)$ were 0.076, 0.082 and 0.064, respectively. These results indicate that all three models achieve a good fit with respect to the savings data (because they each correspond to a high p-value in the K-S hypothesis test), with the logarithmic function slightly outperforming the other two specifications. Because the logarithmic function achieves a better fit and also produces a higher estimate of Medicare's savings in the simulation,

Table EC.1 Bootstrap estimates for subsidy-based contract (OC-I) under alternate minimum savings thresholds, in millions.

MSR	Medicare Savings
0.00%	\$190 (\$107, \$271)
0.25%	\$196 (\$117, \$283)
0.50%	\$197 (\$114, \$293)
0.75%	\$202 (\$99, \$313)
1.00%	\$196 (\$115, \$310)
1.25%	\$198 (\$98, \$304)
1.50%	\$204 (\$133, \$292)
1.75%	\$210 (\$117, \$329)
2.00%	\$207 (\$97, \$328)
2.25%	\$196 (\$97, \$290)
2.50%	\$189 (\$106, \$272)
2.75%	\$190 (\$91, \$305)
3.00%	\$172 (\$77, \$274)
3.25%	\$184 (\$56, \$267)
3.50%	\$160 (\$61, \$241)
3.75%	\$151 (\$52, \$242)
4.00%	\$119 (\$27, \$190)

these results suggest that the estimates produced using $c(x,\theta) = x^2/\theta$ may be conservative, and that Medicare's savings under the optimal contract may in fact be higher than what is reported in Section 7.

Table EC.2 Bootstrap estimates for subsidy-based contract (OC-I) under alternate investment functions, in millions.

		Baseline		Optimized		Δ		
$c(x,\theta)$		Mean	95% C.I.	Mean	95% C.I.	Mean	95% C.I.	p
x^2/θ	ACOs	\$282	(\$188, \$382)	\$316	(\$216, \$423)	\$34	(\$22, \$49)	<0.01
	Medicare	\$146	(\$39, \$260)	\$207	(\$97, \$328)	\$62	(\$48, \$79)	<0.01
	Total	\$427	(\$234, \$642)	\$523	(\$320, \$741)	\$96	(\$78, \$118)	<0.01
$(x^2+x)/\theta$	ACOs	\$256	(\$162,\$353)	\$284	(\$188, \$389)	\$28	(\$18, \$42)	<0.01
	Medicare	\$137	(\$31, \$249)	\$187	(\$79, \$303)	\$50	(\$38, \$63)	<0.01
	Total	\$393	(\$200, \$598)	\$449	(\$250, \$660)	\$78	(\$63, \$94)	<0.01
$(x/\theta)\log(x+1)$	ACOs	\$221	(\$127, \$326)	\$249	(\$137, \$355)	\$27	(\$10, \$47)	<0.01
	Medicare	\$135	(\$8, \$272)	\$383	(\$228, \$535)	\$248	(\$194, \$321)	<0.01
	Total	\$356	(\$150, \$591)	\$632	(\$392, \$877)	\$276	(\$214, \$354)	<0.01

Noisy benchmark observations. In general, we have assumed that the benchmark μ represents the ACO's expected spending on healthcare delivery in the status-quo, that is, with no ACO

investment. In practice, the ACO's expected spending under no investment and the benchmark assigned by Medicare may be different. We numerically consider a setting where the true benchmark for a given ACO is μ, which is known to the ACO, but Medicare observes it to be $\mu + \kappa$, where κ is a random noise term. We set κ to be a zero-mean uniform random variable with support $[-0.05\mu, 0.05\mu]$. In other words, we consider a setting where Medicare's observation of each ACO's benchmark may be up to 5% higher or lower than the true benchmark. The results in Table EC.3 indicate that inaccurate estimates of the benchmark can decrease Medicare's savings under both the baseline and subsidy-based contracts. However, the improvement potential associated with the performance-based subsidy remains fairly constant at \$60 million in both cases. The ACO's total payoff increases when Medicare's observation of the benchmark is noisy, which is unsurprising given the informational advantage that ACOs enjoy in this setting.

Benchmark-based contract. Lastly, we consider a contract that depends only on the ACO's benchmark, instead of its type parameter. This contract may be considered be more practical than a menu of contracts over ACO types, given that the benchmark is a tangible and observable attribute of each ACO. Specifically, we consider a contract that assigns a single shared savings rate α and subsidy rate β to each of the three benchmark clusters from Section 7. Further, we assume in this setting that Medicare observes a noisy signal of the ACO's investment. We posit that (partially) observing the ACO's investment can be feasible in practice since Medicare already extensively monitors ACOs within the MSSP, and has a legislated mandate to audit Medicare providers in general. To reflect the incentive that ACOs have to inflate their reported investments (so to earn a larger subsidy), we assume Medicare observes the ACO's investment to be $(1+\eta)c(x,\theta)$, where η is an exponential random variable with a mean of 0.1. In other words, we assume Medicare observation of the true investment is always inflated, with an average inflation of 10% above the true investment.

Figure EC.1 shows the optimal contract parameters for the benchmark-based contract. The optimal shared savings rate for all benchmark groups was found to be 0.55. This is a consequence of the the contract being unable to distinguish between ACO types combined with Proposition 1. Note also that the subsidy rate is $\beta = 0$ for low benchmark ACOs, since low benchmark ACOs are all concentrated as low-type ACOs, and thus ineffective at generating savings (cf. Table 3). As a consequence, it is optimal for Medicare to not offer the investment subsidy to low benchmark ACOs. By contrast, the optimal subsidy rate for intermediate and high benchmark ACOs is $\beta = 0.3$. Table EC.4 shows the simulated savings under the regular and benchmark-based contracts. Interestingly, the benchmark-based contract achieves nearly the same level of Medicare savings as the full type-based contract (\$187 vs \$207). This is likely due to the fact that the optimal contract parameters in the type-based contract are relatively insensitive to the ACO type, as shown in Figure 5b).

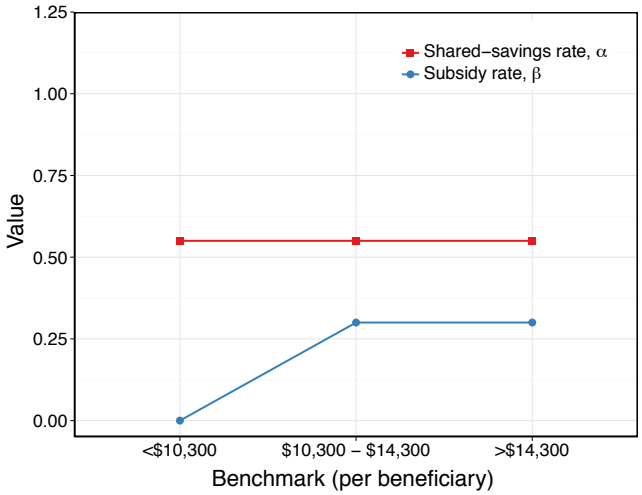


Figure EC.1 Optimal contract parameters in benchmark-only contract.

Table EC.3 Bootstrap estimates for subsidy-based contract (OC-I) under noisy benchmark observations, in millions.

		Baseline		Optimized		Δ		
μ-Error		Mean	95% C.I.	Mean	95% C.I.	Mean	95% C.I.	p
0%	ACOs	\$282	(\$188, \$382)	\$316	(\$216, \$423)	\$34	(\$22, \$49)	<0.01
	Medicare	\$146	(\$39, \$260)	\$207	(\$97, \$328)	\$62	(\$48, \$79)	<0.01
	Total	\$427	(\$234, \$642)	\$523	(\$320, \$741)	\$96	(\$78, \$118)	<0.01
5%	ACO	\$309	(\$194, \$421)	\$337	(\$218, \$452)	\$27	(\$15, \$41)	<0.01
	Medicare	\$111	(−\$26 , \$219)	\$172	(\$31, \$292)	\$60	(\$44, \$82)	<0.01
	Total	\$420	(\$169, \$639)	\$508	(\$251, \$723)	\$88	(\$68, \$111)	<0.01

Table EC.4 Bootstrap estimates for subsidy-based contract (OC-I) with partial observability of ACO investment, in millions).

		Baseline		Optimized		Δ		
Contract		Mean	95% C.I.	Mean	95% C.I.	Mean	95% C.I.	p
θ-Based	ACOs	\$282	(\$188, \$382)	\$316	(\$216, \$423)	\$34	(\$22, \$49)	<0.01
	Medicare	\$146	(\$39, \$260)	\$207	(\$97, \$328)	\$62	(\$48, \$79)	<0.01
	Total	\$427	(\$234, \$642)	\$523	(\$320, \$741)	\$96	(\$78, \$118)	<0.01
μ-Based	ACOs	\$278	(\$182, \$375)	\$332	(\$232, \$437)	\$54	(\$40, \$73)	<0.01
	Medicare	\$127	(\$17, \$233)	\$187	(\$72, \$301)	\$60	(\$45, \$79)	<0.01
	Total	\$406	(\$202, \$602)	\$519	(\$308, \$734)	\$124	(\$99, \$158)	<0.01

EC.4. ACO Savings over Multiple Periods

From the perspective of Medicare, it may be ideal for the ACO to behave myopically, that is, generate as large of a savings as possible each year. However, since the existing MSSP benchmarking

mechanism lowers the benchmark based on the ACO's savings in the previous year, an ACO might find it optimal to strategically curtail investment (and generate lower savings) in earlier years, in order to receive more favorable benchmarks in later years. In this section, we address the multiple-period setting where the ACO may behave in this strategic manner. We briefly discuss relevant literature and potential approaches to modifying the benchmarking methodology to mitigate the investment distortions that might arise from such strategic behavior. We also present a result that shows that for a forward looking ACO, the deviation in savings from the myopic strategy vanishes over time under the existing benchmarking mechanism.

The notion that an agent might find it optimal to strategically delay effort in multi-period incentive problems is well documented in the operations literature. Chen (2000) consider a salesforce compensation contract where agents receive payments based on an annual sales quota system. The author shows how the "sales-hockey stick" phenomenon, where agent effort is delayed to the final period of the horizon, can arise in this setting. Sohoni et al. (2010) also consider threshold-based contracts, and show that adjusting the sales threshold based on a relevant market signal can mitigate the sales-hockey stick phenomenon. In a similar vein, Besbes et al. (2016) consider how the presence of debt in a dynamic pricing setting can lead to price distortions and efficiency losses.

One approach to regulating the strategic underinvestment of a forward-looking ACO might be to set the financial benchmarks of an ACO according to the spending patterns of comparable ACOs, rather than setting benchmarks for each ACO individually based on their historical spending. In a seminal paper, Shleifer (1985) discusses how regulated local monopolies have little incentive to reduce costs in an environment where the regulator sets prices according to the firm's costs. The author proposes a mechanism known as *yardstick competition* whereby the regulator sets prices for a firm according to the costs of identical firms. This mechanism has the effect of inducing competition between firms that serve different markets. With respect to the MSSP, incorporating elements of yardstick competition when setting the financial benchmarks of ACOs may be useful in disincentivizing strategic underinvestment. However, while a purely yardstick competition approach to setting the ACO benchmarks might dampen underinvestment, due to the voluntary nature of the MSSP contract it may also jeopardize ACO participation if some ACOs find the benchmarks to be unattainable, which is an undesirable outcome for Medicare. We therefore posit that a benchmarking methodology that is based on a blend of the ACO's own historical spending and yardstick competition might be a more fruitful approach to mitigating strategic delay of investment, while also encouraging ACO participation, and underline this as a direction for future work. Yardstick competition for controlling costs is not new to Medicare. For example, the Medicare Prospective Payment System (PPS) reimburses providers based on the costs of comparable hospitals (Fetter 1991). This program was created precisely in response to the adverse incentives that arise from

purely cost-based reimbursement systems. For a more recent example, Savva et al. (2016) consider a modification of traditional cost-based yardstick competition in the context of a queueing system, motivated by the problem of reducing wait times in emergency departments.

It is also worth noting that like several other innovations in Medicare provider payment systems, the MSSP was created to realign incentives and reduce provider spending. In contrast to other Medicare programs, however (e.g., the PPS), the MSSP emphasizes self-improvement of each ACO independently, rather than competition between ACOs. While the existing MSSP benchmarking mechanism may allow ACOs to strategically delay investment, and can potentially be improved upon, our analysis suggests that a revised contract with performance-based subsidies can still lead to overall cost improvements. In the remainder of this section, we consider a simple multi-period model for the MSSP in which the ACO's benchmark is updated based on its savings in the previous year. We find that the ACOs distortion in its optimal savings level relative to a myopic policy diminishes over time.

Consider a single ACO with type θ that faces a multiple-period savings problem with an infinite horizon, where $t = 0, 1, 2, \ldots$, indexes each period. Let α and β be fixed throughout, which reflects the fact that the MSSP uses the same shared savings formula each year. We suppress θ, α and β in the notation. Let $\mu_0, \mu_1, \mu_2, \ldots$, be the benchmarks in each period, and let $x_0, x_1, x_2, \ldots$, be the ACO's decision variables, which represents the ACO's reduction in spending. We emphasize that x_t represents the current-period savings relative to the relevant benchmark μ_t. Define $s_t = \sum_{i=0}^{t} x_i$, so that $s_0, s_1, s_2, \ldots$, is the cumulative savings up to and including period t. We assume that the cost of generating savings x_t in period t is $c(s_{t-1} + x_t)$, where $c(\cdot)$ is the single period cost function described in Assumption 1. In other words, the investment required in period t to reduce spending down to a level of $\mu_t - x_t$ is measured with respect to the expected spending in the baseline "do-nothing" case, which is given by μ_0. We note here that if the investment required to achieve a savings of x_t were $c(x_t)$, then it can be shown that the ACO would simply behave myopically each year.

In the MSSP, the benchmarks are updated based on the previous period's benchmark and the savings generated by the ACO. The benchmark only decreases if the ACO generates positive savings, i.e., the benchmark is never increased when ACO spending exceeds it (Federal Register 2011). To reflect the MSSP's current practice, for $t \geq 2$ let $\mu_t = \mu_{t-1} - x_{t-1}$. The initial benchmark, $\mu_0 = \mu$, is determined based on the ACO's historical spending (Federal Register 2011). Note that s_t is then given by $s_t = \mu_0 - \mu_t + x_t$. Next, let $y_t = x_t + \varepsilon$ be the realized savings, as in the single period model. The ACO's total expected payment in period t is then given by $P_t(x_t) = \int_{-\infty}^{\infty} [r(y_t, \alpha) + s(y_t, \beta)]\, \omega(y_t|x_t) dy_t$, and the ACO's profit in period t is then $u_t(x_t) = P_t(x_t) - \gamma x_t -$

$c(s_{t-1}+x_t)$. Let $\delta<1$ be the discounting factor. The ACO's multiple-period optimization problem is then given by

$$\begin{aligned}
\underset{\mathbf{x}}{\text{maximize}} \;& \sum_{t=0}^{\infty} \delta^t \left(P_t(x_t) - \gamma x_t - c(s_{t-1}+x_t)\right) \\
\text{subject to } & s_t = \mu_0 - \mu_t + x_t, \quad t=1,2,\ldots, \\
& \mu_t = \mu_{t-1} - x_t, \quad t=1,2,\ldots, \\
& \mu_0 = \mu, \\
& s_0 = 0, \\
& x_t \in [0,\bar{x}], \quad t=0,1,2,\ldots,
\end{aligned}$$

Let the optimal solution to the above problem be $(\tilde{x}_0,\tilde{x}_1,\tilde{x}_2,\ldots,)$. If the ACO were to reduce spending myopically, then in each period t it would simply solve the single period problem

$$\begin{aligned}
\underset{x_t}{\text{maximize}} \;& P_t(x_t) - \gamma x_t - c(s_{t-1}+x_t) \\
\text{subject to } & x_t \in [0,\bar{x}],
\end{aligned}$$

where $s_{t-1}=\sum_{i=0}^{t-1} x_i$. Let the optimal solution to the myopic problem be $(x_0^*,x_1^*,x_2^*,\ldots)$. We can now interpret $|\tilde{x}_t - x_t^*|$ as the distortion due to the ACOs strategic saving. The following result shows that the distortion

Proposition EC.1. *In the multiple-period setting, the difference in ACO savings under myopic and forward-looking behavior vanishes,* $\lim_{t\to\infty}|\tilde{x}_t - x_t^*| = 0$.

Proof. It suffices to show that for any $\epsilon>0$, there exists $\bar{T}>0$ such that $|\tilde{x}_t - x_t^*|<\epsilon$ for all $t\geq\bar{T}$. First we show that for any ϵ, there exists $\tilde{T}$ such that $\tilde{x}_t\leq\epsilon$ for all $t\geq\tilde{T}$. Suppose there exists $\epsilon>0$ and a subsequence $\tilde{x}_{t_k}$ such that $\tilde{x}_{t_k}>\epsilon$ for all $k\geq 0$. Let $k(t)=\sup\{k\geq 0|t_k\leq t\}$. Since $\tilde{x}_t\geq 0$ for all $t\geq 0$, for any t we have $\sum_{i=0}^{t}\tilde{x}_i\geq\sum_{k=0}^{k(t)}\tilde{x}_{t_k}$. Then we have $\lim_{t\to\infty} s_t = \lim_{t\to\infty}\sum_{i=0}^{t}\tilde{x}_i \geq \lim_{t\to\infty}\sum_{k=0}^{k(t)}\tilde{x}_{t_k} \geq \lim_{t\to\infty}\sum_{k=0}^{k(t)}\epsilon=\infty$. Since $c(s_t)$ is strictly increasing by Assumption 1, it follows that $\lim_{t\to\infty} u_t(\tilde{x}_t) = -\infty$, which cannot be optimal. Therefore, there must exist $\tilde{T}>0$ such that $\tilde{x}_t\leq\epsilon$ for all $t\geq\tilde{T}$. By a parallel argument, for any $\epsilon>0$ there exists T^* such that $x_t^*\leq\epsilon$ for all $t\geq T^*$. The result follows by taking $\bar{T}=\max\{T^*,\tilde{T}\}$. $\square$

Proposition EC.1 is a straightforward result – it shows that the deviation from the myopic behavior in a multi-period setting must vanish because the ACO savings itself under both myopic and forward-looking behavior goes to 0. This result follows from the fact that the monotonic reduction in the financial benchmarks within the MSSP makes it increasing costly for ACOs to receive bonus payments. Therefore, while the existing benchmarking mechanism may allow ACOs to strategically delay investment, this behavior is unlikely to persist in the long term.

EC.5. Bootstrap Simulation

Here we outline the steps used to simulate MSSP performance under the optimal contract. Let α_k^* and β_k^* be the optimal contract parameters corresponding to each $\theta_k \in \Theta$. The steps for producing the bootstrap samples for the first-best contracts are given in Algorithm 1 below. Each of the L bootstrap samples is itself constructed by repeatedly sampling (with replacement) from the ACO data a total of n times, which is a standard bootstrapping practice (Efron and Tibshirani 1994). Note that in each iteration, we multiply the simulated savings v_s' by the number of beneficiaries, b_s, so that a single sample v_ℓ and u_ℓ represents the total, non-normalized Medicare savings (similarly for the total ACO payoff).

Algorithm 1 Bootstrap procedure for simulating MSSP performance

Input: Data: $(\mu_i, b_i, y_i, \theta_i)$, $i = 1, \ldots, n$. Optimal ACO savings: $x(\alpha, \beta, \theta)$ for all $\alpha \in \mathcal{A}$, $\beta \in \mathcal{B}$, $\theta \in \Theta$. Optimal contract parameters: α_k^* and β_k^* for $\theta_k \in \Theta$. Benchmark groups $\mathcal{M}_1, \ldots, \mathcal{M}_m$. Estimated parameters: $\hat{\sigma}$ and $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$.

for $\ell = 1, \ldots, L$ **do**

  for $s = 1, \ldots, n$ **do**

    Sample (μ_s, b_s) from $(\mu_1, b_1) \ldots, (\mu_n, b_n)$.

    Sample θ_s from $f(\theta|\hat{\lambda}_j)$, for j such that $\mu_s \in \mathcal{M}_j$.

    $\alpha_s \leftarrow \alpha_j^*$, $\beta_s \leftarrow \beta_j^*$.

    $x_s \leftarrow x(\alpha_s, \beta_s, \theta_s)$.

    Sample ξ_s from $g(\xi|\hat{\sigma})$.

    $y_s \leftarrow x_s + \xi_s$.

    $v_s' \leftarrow x_s - r(y_s, \alpha_s) - s(y_s, \beta_s)$.

    $u_s' \leftarrow r(y_s, \alpha) + s(y_s, \beta) - c(x, \theta_s) - \gamma x_s$.

  $v_\ell \leftarrow \sum_{s=1}^n v_s' b_s$.

  $u_\ell \leftarrow \sum_{s=1}^n u_s' b_s$.

Output: Bootstrap samples $v_1, \ldots, v_L$, and $u_1, \ldots, u_L$.

References

Besbes, O., Iancu, D. A., & Trichakis, N. (2017). Dynamic pricing under debt: Spiraling distortions and efficiency losses. *Management Science*.

Chen, Fangruo. 2000. Sales-force incentives and inventory management. *Manufacturing & Service Operations Management* **2**(2) 186-202.

Fetter, Robert B. 1991. Diagnosis related groups: understanding hospital performance. *Interfaces* **21**(1) 6-26.

Savva, Nicos, Tolga Tezcan, Ozlem Yildiz. 2018. Yardstick competition for service systems. *Management Science.*

Shleifer, Andrei. 1985. A theory of yardstick competition. *The RAND Journal of Economics* 319-327.

Sohoni, Milind G, Achal Bassamboo, Sunil Chopra, Usha Mohan, Nuri Sendil. 2010. Threshold incentives over multiple periods and the sales hockey stick phenomenon. *Naval Research Logistics* (NRL) **57**(6) 503-518.