Design of Medical Reimbursement Policy and Effects of Pooling

Yiwen Shen
School of Business and Management, Hong Kong University of Science and Technology, yiwenshen@ust.hk
Qingfa Zhang
School of Business and Management, Hong Kong University of Science and Technology, qzhangcl@connect.ust.hk
Ying-Ju Chen
School of Business and Management, Hong Kong University of Science and Technology, imchen@ust.hk

Abstract. We study the design of medical reimbursement policy using a model with multiple types of agents and services. The central planner decides the reimbursement policy for agents' medical costs to minimize the total utility loss, subject to a budget constraint. We find that the optimal reimbursement policy has a cap structure, in which all excess costs above a threshold are reimbursed. We also solve the optimal structure for a ratio reimbursement policy, which is commonly used in practice. Then, we investigate the effects of horizontal pooling when there are multiple groups. We show that with the optimal cap policy or under homogeneous cost distributions, it is optimal to integrate the groups together and use a common reimbursement policy for all. Finally, we consider a dynamic model in which the central planner needs to balance the current and future utilities. In the optimal dynamic policy, the central planner should increase the marginal spending rate for current period when the reimbursement scope is enlarged, leading to a piecewise concave structure of optimal spending amount. An extensive numerical study shows the value of horizontal pooling and dynamic policy in medical insurance.

Key words: Medical services, reimbursement policy, utility optimization, pooling, dynamic programming

1. Introduction

In many countries, the public sector (e.g., government as the single-payer) plays a major role in developing and running the healthcare insurance systems for their population (e.g., mainland China, U.K., Canada, and Sweden). In such systems, an agent's medical cost is reimbursed by the insurance scheme in which she participates according to a prespecified policy. In recent decades, healthcare expenditures have been driven up by multiple factors, including aging populations, technology advancements, and prevalence of chronic diseases (WHO 2019). Thus, providing efficient medical services while ensuring the financial sustainability has become a critical challenge for many healthcare systems in the world. Under the constraint of limited budget, the policy makers need to decide how to allocate the insurance funds across different demographic groups (e.g., rural and urban residents) and medical services (e.g., inpatient and outpatient). This resource allocation problem is reflected by the design of reimbursement policy, which determines the reimbursement amount for a given cost level when an agent requests a medical service.

In modern societies, providing equitable and universal healthcare access and service has been a commonly adopted goal (Atun et al. 2015, Cotlear et al. 2015). However, in many places, the medical insurance systems still operate in a separate and fragmented way. For example, in mainland China, different insurance

schemes are offered based on the employment and residency status of citizens. The fragmentation of insurance systems leads to substantial imbalance in the expenditure levels, reimbursement policies, and benefit packages of different groups. This significantly hinders the healthcare access and financial protection for the groups with lower social-economic rankings, leading to social welfare loss and violation of the fairness principle (Culyer and Newhouse 2000). To overcome this challenge, integration and consolidation of medical insurance schemes have been implemented in different countries (Kwon 2008, Ikegami et al. 2011), although more efforts are still needed. Economically, this is akin to the operational strategy of resource pooling, which has been widely used in different settings of operations management (Benjaafar 1995, Corbett and Rajaram 2006).

In this paper, we study the design of medical reimbursement policy from the central planner's aspect. In our base model, we consider the reimbursement problem for a group of agents in a single period. There are two types of medical services, with the second-tier service being on average more costly than the first-tier one. Each agent has an exogenous probability to require one of the two types of services, and therefore incurs a random service cost. The central planner decides the reimbursement policy for the two services, i.e., the reimbursement amount as a function of the actual cost and service type. For the central planner, there is a utility loss associated with the agent's net medical cost after reimbursement. We assume the central planner is risk averse and prefers to reduce the extreme costs of agents (Culyer and Newhouse 2000). The central planner minimizes the expected total utility loss of all agents in the group, subject to a budget level constraint. We formulate this as an optimization problem over the function space and solve it explicitly based on the variation calculus method (Luenberger 1997).

We find that the optimal reimbursement policy has a max-out-of-pocket structure. That is, the central planner reimburses all excess medical costs above a common self-payment threshold. This is also referred to as the cap policy, as agent's net cost is capped from above. Intuitively, the cap policy is optimal as it is most efficient in avoiding large net costs of agents. In addition, we also solve the optimal ratio reimbursement policy, in which the central planner reimburses a fixed proportion of agents' medical costs depending on the service type. It has the following structure. When the budget is low, the central planner only reimburses the service with higher and more dispersed cost. As the budget increases, both types of services are reimbursed. In addition, the marginal benefit from investing in the two services coincide, suggesting the central planner cannot improve the objective by redistributing her budget across the two services. The cap and ratio reimbursement policies, as well as their mixes, are widely used in reality by different countries.

We extend our model to multi-group setting to study the effects of horizontal pooling. In this setting, the central planner manages multiple groups with potentially different population size, budget level, service incidence, and cost distributions. We consider three pooling systems differentiated by two operational factors: resource sharing and policy flexibility. In the non-pooling system, the groups operate independently with their own budget and policy. In the full pooling system, the groups are fully combined into one with a

shared budget and a common reimbursement policy for all. In the monetary pooling system, the groups still share their budgets, but different reimbursement policies can be used for each group. We solve the optimal cap and ratio policies explicitly in the three pooling systems, similar to those in our single-group model.

The monetary pooling always leads to the best performance as it has both resource sharing and policy flexibility. However, two practically relevant questions remain regarding the performance of the pooling systems. The first is under what circumstances would a common reimbursement policy be optimal for all groups. The second is whether the central planner should choose policy flexibility or resource pooling when facing such trade-off. We develop analytical answers for the two questions. We find that when (i) the optimal cap policy is used or (ii) the optimal ratio policy is used with homogeneous cost distributions across groups, the full pooling system achieves the same performance as monetary pooling, and thus outperforms the non-pooling system. In these cases, it is optimal to use a common reimbursement policy for all groups, suggesting that policy flexibility is not needed. However, when optimal ratio policy is used with heterogeneous cost distributions, full pooling system may underperform the non-pooling system. In this case, the benefit of policy flexibility may outweigh that of resource sharing. In reality, heterogeneous costs can occur when one group is on average less healthy than the other (e.g., with more senior adults or chronic diseases). These findings provide managerial insights when consolidating the medical insurance schemes and designing reimbursement policies for multiple groups.

Next, we develop our model in a dynamic setting to capture the intertemporal trade-off in the central planner's decision-making. The central planner manages a medical fund to minimize the total discounted utility loss over all future periods. In each period, she decides the spending amount for current period and the corresponding reimbursement policies. The remaining fund earns an interest rate and is carried to next period. In addition, there is a random inflow to the fund in each period. We formulate the problem as a Markov decision process (MDP) and characterize the optimal solution. We find that under certain conditions, the per period optimal spending amount has a piecewise concave structure with respect to current fund level. In particular, there is an upward jump in the marginal spending rate when a new service is included in the reimbursement scope. This can be interpreted as enlarging the reimbursement scope slows down the decline rate in the marginal benefit of spending. Finally, we show that the main results regarding the performance of the three pooling systems hold under the dynamic setting.

We conduct multiple numerical experiments based on our model solution. Our results show that the improvement from horizontal pooling is larger when there is higher imbalance between the groups. On the other hand, the benefit of using the dynamic policy (over the myopic policy) is greater when the interest rate is higher and when the inflow is more volatile. We further use a counterfactual study to demonstrate the benefit of pooling based on real-world data. In mainland China, two main medical insurance schemes are used for urban employees and rural & urban residents, respectively. However, there is large imbalance in the expenditure level of the two schemes, with about four times difference in the per capita spending (Meng

et al. 2015). We evaluate the hypothetical benefit of fully integrating the two schemes for two gynecological conditions: uterine myoma and cesarean section. Using real-world health statistics, we show that pooling can lead to substantial economic benefit for the central planner, as it reduces the large imbalance between the two schemes. Such implication is in line with the current health reform in mainland China, where the pilot integration of the two schemes has started in some more developed cities (Pei 2014).

In this work, we focus on the resource allocation aspect in medical insurance and make two main contributions. First, we establish an analytical framework for the central planner's decision-making in both multi-group and dynamic settings. We explicitly solve the structure of optimal policy, which can be useful for policy makers. Second, we investigate the effects of pooling with budget sharing and/or policy flexibility. We develop sufficient conditions for the optimality of using a common reimbursement policy for all groups in both single-period and dynamic models. The results provide managerial insights on the trade-off between resource sharing and policy flexibility in the multi-group settings, and are practically relevant for the integration and consolidation of different insurance schemes. The benefit of pooling is further demonstrated by numerical experiments and a counterfactual study based on real-world data.

The rest of the paper is organized as follows. In the next section, we provide a brief review of related literature. Section 2 introduces the basic model setup and solves the optimal reimbursement policies. Section 3 extends the model to multi-group setting and analyzes the effects of horizontal pooling. In Section 4, we develop a dynamic model for the reimbursement problem using a Markov decision process. Section 5 conducts an extensive numerical study. Section 6 concludes and discusses future directions. Auxillary results and proofs are collected in the electronic companion.

1.1. Literature Reviews

Our work is related to the following streams of literature: social welfare implications of medical insurance, design of reimbursement and payment policy, resource pooling, and integration of health insurance systems.

The importance of medical insurance has been widely recognized in modern societies. The seminal work of Arrow (1963) lays a theoretical foundation. It shows that health insurance can improve social welfare via risk sharing in the presence of uncertainty. Subsequent works further investigate the implications of health insurance in different settings. In general, the expansion of healthcare insurance is associated with positive outcomes, such as lower health inequality (Wang et al. 2009), enhanced financial security and protection (Gross and Notowidigdo 2011), increased healthcare utilization, and better health outcomes (Finkelstein et al. 2012, Sommers et al. 2017). For such benefits, providing equitable and universal healthcare access and service for the entire society has been a commonly adopted goal by policy makers (Atun et al. 2015, Cotlear et al. 2015). However, the budget for health insurance is usually limited in practice, and the medical costs have been driven up by multiple factors such as aging population and chronic diseases. This calls for designing more efficient and fair medical insurance systems to better allocate the limited resources among participants (Baicker and Goldman 2011).

One key element in medical insurance scheme is the design of reimbursement and payment policies, which determine how much a person has to pay out of pocket for his or her medical expenses. In addition, these policies may affect both the demand and supply sides of medical services (Clemens and Gottlieb 2014, Finkelstein et al. 2019). The design of insurance policy has been studied in health economics from multiple aspects, such as consumer risk preferences (Einav et al. 2010), adverse selection (Handel et al. 2015), and provider-insurer bargaining (Ho and Lee 2019). Some recent works in operations management focus on the behaviors and decision-making of healthcare providers under different payment systems. For example, Adida et al. (2017) and Guo et al. (2019) study the payment contracts for coordinating the capacity and quality decision of healthcare providers. Dai et al. (2017) investigate the cream-skimming behavior of healthcare providers under different reimbursement schemes. Different from these works, we study the reimbursement policy design from a utility optimization aspect, where the central planner minimizes the total utility loss associated with agents' net costs. This can be viewed as a resource allocation problem — the reimbursement policy determines how the limited resource (budget) is allocated to agents with different service types and cost levels.

We investigate the effects of pooling when the central planner manages the medical insurance of multiple groups. This is related to the literature on resource pooling, which is extensively studied in operations management. Eppen (1979) shows that pooling demands and inventory in a multi-location newsvendor setting reduces the total costs (see also Yang et al. 2021). As an important operational strategy, resource pooling has been applied in multiple business areas, such as production systems (Benjaafar 1995), revenue management (Afeche and Pavlin 2016), supply chain (Qi et al. 2015), inventory management (Corbett and Rajaram 2006). In healthcare operations, different forms of pooling have been considered, such as sharing medical equipment between hospitals (Deo and Gurvich 2011), pooling servers in queuing systems (Mandelbaum et al. 2012), diverting patients to less crowded hospitals (Xu and Chan 2016), and sharing capacity across surgery specialties (Song et al. 2020). We study the effects of pooling at the macro level in the medical reimbursement problem, and develop analytical results for the value of budget sharing and policy flexibility. This contributes to the literature by revealing the potential benefits of pooling in social policy-makings.

From the practical aspect, our study is related to the integration and consolidation of medical insurance schemes. In many regions, the healthcare systems still operate in a fragmented way, with substantial heterogeneity in the financial protection, service access, and benefit package for different socio-economic groups (Lagomarsino et al. 2012, Atun et al. 2015). This is because medical insurance is usually developed in a sequential way: first targeting groups with higher ability to pay (e.g., large corporation employees) and then expanding to lower-income groups (e.g., rural residents). Various efforts have been made by policy makers to promote the integration of different medical insurance systems. Examples include South Korea (Kwon 2008), Japan (Ikegami et al. 2011), and mainland China (Meng et al. 2015) among others. Empirical studies

find that the integration of medical insurance schemes leads to improved health outcomes and decreased inequality (Yang et al. 2018). Different from these works, we develop a rigorous utility-optimization framework for solving the optimal reimbursement policy under different pooling systems. Our results provide analytical support for the benefits of integrating different medical insurance schemes.

2. Single-Group Model

In this section, we develop our base model for studying the optimal reimbursement problem in medical services. The model serves as a foundation for our analysis in subsequent sections.

2.1. Model Set-up

We consider a group of agents with a continuous mass of N. The population size is assumed to be large (e.g., all employees in a city), as for most medical insurance schemes in reality. The agents are indexed by $\iota \in [0, N]$. There are two types of medical services: the first-tier service (F) and the second-tier service (S). The second-tier service is on average more complex and costly for the agents. Our results can be easily extended to more types of services. The central planner reimburses agents' medical costs based on the service type. That is, one common reimbursement policy is used for costs associated with each service type, but policies can vary for the two types of services. The definitions of service types depend on the medical setting and granularity of reimbursement policy. For example, the first and second-tier services can represent general outpatient and inpatient care, respectively. For a given disease, the two service types can refer to treatments for patients with low and high severity levels (e.g., early and advanced stage cancers).

We consider the central planner's decision-making in a given period. The period length is defined such that it is reasonable to assume an agent requires at most one service in a period. For agent ι, she has a probability $p_{\iota F}$ to require the first-tier service, probability $p_{\iota S}$ to require the second-tier service, and probability $1 - p_{\iota F} - p_{\iota S}$ to require no service. If service of type $j \in \{F, S\}$ is required, the agent ι incurs a random cost of $C_{\iota j}$, which follows cumulative distribution function (CDF hereafter) $G_{\iota F}(\cdot)$ or $G_{\iota S}(\cdot)$. We assume the service costs are always positive with common supports $(\underline{C}_F, \overline{C}_F)$ and $(\underline{C}_S, \overline{C}_S)$ for the two types, respectively.[1] We allow the service probabilities and cost distributions to vary across agents to reflect the heterogeneity in their health conditions and demands.

The central planner decides the reimbursement policy for the agents' medical costs. The reimbursement policy is described by two functions $\phi_F(x)$ and $\phi_S(x)$, which represent the reimbursement amount for service F and S if the raw cost is x. We allow the reimbursement policy to vary for the two types of services, as shown by the subscript F and S. We require the reimbursement policy to satisfy $0 \le \phi_F(x), \phi_S(x) \le x$. That is, the reimbursement amount should be non-negative but not exceed the raw cost. As a natural

[1] The rare diseases with potentially unbounded costs are out of scope of this work.

assumption, we only consider continuous functions $\phi_F(x)$ and $\phi_S(x)$. We define the admissible set for reimbursement policy as

$$\mathcal{C} := \{(\phi_F(\cdot), \phi_S(\cdot)) : \phi_F(x) \text{ and } \phi_S(x) \text{ are continuous with } 0 \leq \phi_F(x), \phi_S(x) \leq x \text{ for all } x\}. \quad (1)$$

The central planner faces a budget constraint when designing the reimbursement policy. In particular, the expected total amount of reimbursement for both types of services cannot exceed its budget. We measure the budget level by the per capita funding amount, denoted by m. Thus, the central planner's total budget is given by Nm. The budget for reimbursement includes both agents' medical insurance payments to the central planner and potential funding from other sources (e.g., transfer payments).

With reimbursement policy $\phi_j(x)$ for $j \in \{F, S\}$, an agent with raw cost x incurs a net (out-of-pocket) cost of

$$l_j(x) = x - \phi_j(x), \quad j \in \{F, S\}. \quad (2)$$

The central planner minimizes the expected total utility loss associated with agents' net costs. Denote the utility loss by $u(l)$ for $l > 0$. We interpret $u(l)$ as the utility loss of a representative agent, which is commonly used in the macroeconomics literature (Hartley 1997). We assume $u(l)$ strictly convex increases in l with continuous second-order derivative. In addition, it satisfies $u(0) = u'(0) = 0$. The convexity of $u(l)$ implies that $\mathsf{E}[u(L)] \geq u(\mathsf{E}[L])$ for a positive random cost L. This reflects the risk aversion of central planner when designing reimbursement policies. In particular, the central planner aims to avoid extreme costs for agents, which is the fundamental motivation and commonly accepted goal of medical insurance (see, e.g., Arrow 1963). The actual utility loss is bounded as we assume finite supports of service costs.

A representative example satisfying the above assumptions is the power utility loss function, given by

$$u(l) = \frac{1}{\gamma} \times l^{\gamma}, \quad \text{with } \gamma > 1. \quad (3)$$

Here γ is the risk aversion coefficient. A larger γ implies the central planner is more risk averse. For its analytical convenience, the power utility function in the form of (3) is widely used in economics, finance, and marketing (e.g., Merton 1992, Liu et al. 2018, Apesteguia et al. 2020). In subsequent analysis, we develop our main findings under the general utility loss function $u(l)$. In certain cases, we use the power utility loss function to derive explicit results.

The central planner decides the reimbursement policy to minimize the total utility loss of agents, subject to the budget constraint. We show that the optimization problem can be formulated as follows.

Lemma 1 *The central planner's utility loss minimization problem can be formulated as:*

$$U(m) = \min_{(\phi_F, \phi_S) \in \mathcal{C}} \quad h_F \mathsf{E}[u(l_F(C_F))] + h_S \mathsf{E}[u(l_S(C_S))] \quad (4)$$

$$\textit{s.t.} \qquad h_F \mathsf{E}[\phi_F(C_F)] + h_S \mathsf{E}[\phi_S(C_S)] \le m, \tag{5}$$

where $l_j(x) = x - \phi_j(x)$ for $j \in \{F, S\}$; the aggregated services incidence h_F and h_S are defined as

$$h_j := \frac{1}{N} \int_0^N p_{\iota j} d\iota, \quad j \in \{F, S\}. \tag{6}$$

The aggregated service costs C_F and C_S are random variables with the following CDFs:

$$G_j(x) := \mathsf{Pr}(C_j \le x) = \frac{1}{N h_j} \int_0^N p_{\iota j} G_{\iota j}(x) d\iota, \quad \forall j \in \{F, S\},$$

where $G_{\iota j}(x) = \mathsf{Pr}(C_{\iota j} \le x)$ is the CDF of the individual service cost $C_{\iota j}$.

In the optimization problem (4), the services incidence h_F and h_S are defined as the population average probability that a service is requested in the period. In addition, we introduce two random variables C_F and C_S to represent the aggregated service costs. By (1), their distributions are defined using the CDFs of the individual costs $G_{ij}(x)$, weighted by the service probability $p_{\iota j}$. For the central planner's optimization problem, she only needs to know the population-based services incidence (h_F, h_S) and the distributions of the aggregated service costs (C_F, C_S). Thus, it is sufficient to use (h_F, h_S) and (C_F, C_S) in subsequent analysis, without explicitly involving the individual quantities $p_{\iota j}$ and $C_{\iota j}$. This facilitates the solution of the problem. In particular, h_F and h_S represent the proportions of population that require the services in one period. The distributions of C_F and C_S can be estimated using the empirical distributions of the actual service costs of agents.

The first equation in (4) represents the expected total utility loss, which sums over the two types of services based on the net cost function $l_j(x)$. The second equation represents the budget constraint, i.e., the expected total amount of reimbursement cannot exceed the budget level. As we are considering a relatively large group of agents, the random fluctuations around the expectations are of smaller orders of magnitudes, and are thus ignored in the problem. The full-coverage budget level $\bar{m}$ is given by:

$$\bar{m} := h_F \mathsf{E}[C_F] + h_S \mathsf{E}[C_S]. \tag{7}$$

When $m \ge \bar{m}$, the central planner can reimburse all costs of the agents with $\phi_F(x) = \phi_S(x) = x$. Thus, we only need to consider $m \le \bar{m}$ in our analysis. In this case, we can replace "$\le$" by "$=$" in the budget constraint, as it is always beneficial to spend more budget for reimbursement.

Designing and implementing medical insurance policy is a complex problem that involves many different aspects, such as design of payment systems (fee-for-service versus diagnosis related groups), pricing of new medicines and procedures, resource allocation, and capacity planning of primary and specialty cares (see, e.g., Glied and Smith 2011). In this work, we focus on the resource (budget) allocation problem in medical

insurance by solving the optimal reimbursement policy and revealing the effects of pooling. To facilitate our analysis, we assume the service costs and agents' demands are exogenously given in the model. Thus, our results are mostly applicable to the settings where the service prices are already set and the service demands are stable. Empirical studies find that while patients' demands can respond to insurance policy, the effects are relatively mild compared to the variation in healthcare spending (Brot-Goldberg et al. 2017). In addition, the classification and charging of medical services are usually regulated by healthcare authorities with limited flexibility (Tikkanen et al. 2020). We defer the modeling of endogenous pricing and demand to future research.

2.2. Optimal Reimbursement Policy

In this section, we solve the optimal reimbursement policy for the central planner's optimization problem in (4). We first solve the optimal policy without imposing any restriction on the form of reimbursement function $\phi_j(x)$. This is developed in the following proposition.

Proposition 1 *The optimal reimbursement policy for problem (4) is given by*

$$\phi_j^*(x) = \max\{x - \tau^*, 0\}, \quad \text{for } j \in \{F, S\}. \tag{8}$$

The self-payment threshold τ^* is determined by the budget constraint:

$$\sum_{j\in\{F,S\}} h_j \mathsf{E}[\max\{C_j - \tau^*, 0\}] = m, \tag{9}$$

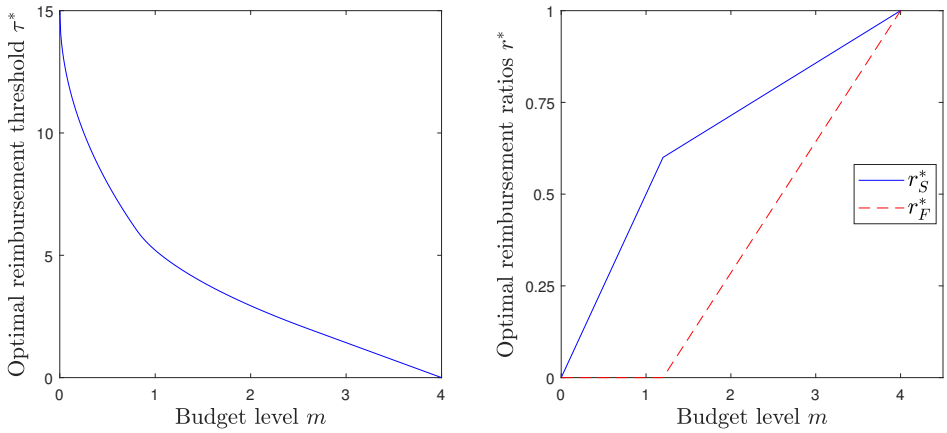
where the expectations are taken over the distributions of costs C_F and C_S. The threshold τ^* convex decreases in m.

Remark 1 *The self-payment threshold τ^* becomes smaller when we replace the random costs C_F and C_S with their expectations $\mathsf{E}[C_F]$ and $\mathsf{E}[C_S]$, i.e., $\tau^*_{random} \geq \tau^*_{constant}$.*

By above proposition, we see that an "max-out-of-pocket" reimbursement policy is optimal to the central planner's problem (4). The central planner fully reimburses the excess part of cost above the threshold τ^*, which is determined by the budget level m. For agents, their net costs are capped at τ^*. Thus, we also refer to the policy in (8) as the *cap* reimbursement policy. The cap structure of the optimal policy stems from the convexity of the utility loss function $u(l)$. For the central planner, it is optimal to first spend her budget on the agents with higher costs, as they are associated with larger marginal utility loss $u'(l)$. Thus, the reimbursement amount aligns with the actual cost, leading to the optimal cap policy.

In the optimal cap policy, the threshold τ^* for self-payment convex decreases in the budget level m. See the left panel of Figure 1 for an illustrative example. Thus, as the budget level increases, its marginal impact

Figure 1 Optimal Reimbursement Policy by Budget Level



on the threshold τ^* becomes smaller. Note that we do not need to distinguish between the first and second-level services in the optimal cap policy, as we always reimburse the agents with higher actual costs first. Finally, we can show that the randomness in the costs C_F and C_S increases the self-payment threshold τ^* compared with the deterministic case (see Remark 1). This reveals the negative impact of random variation in the medical costs.

In the follows, we further consider another type of reimbursement policy, in which the central planner reimburses a fixed proportion of the service costs. We refer to this as the *ratio* reimbursement policy. The reimbursement ratios can vary for the two types of services. Although the ratio policy is not optimal in terms of utility optimization, it is widely used in practice for its convenience. Denote the reimbursement ratios by r_F and r_S, the reimbursement and net cost functions specify to

$$\phi_j(x) = r_j x, \quad l_j(x) = (1 - r_j)x, \quad \forall j \in \{F, S\}. \tag{10}$$

In this case, the central planner only needs to decide $r_F, r_S \in [0, 1]$. Plugging (10) into the formulation (4), we can show the optimization problem is convex in (r_F, r_S), and the optimal solution can be obtained by standard Karush–Kuhn–Tucker conditions. Before presenting the results, we first define a useful quantity related to the service costs.

Definition 1 *Define cost indexes b_F and b_S for the two medical services as:*

$$b_F = \frac{\mathsf{E}[u'(C_F)C_F]}{\mathsf{E}[C_F]} \quad \textit{and} \quad b_S = \frac{\mathsf{E}[u'(C_S)C_S]}{\mathsf{E}[C_S]}. \tag{11}$$

The cost index is solely determined by the cost distribution. Due to the convexity of $u(l)$, $u'(l)$ is monotonically increasing. Thus by (11), the cost index tends to be larger for the service with higher and more dispersed costs. With constant costs c_F and c_S, the cost indexes reduce to the marginal utility loss $b_j = u'(c_j)$ for service $j \in \{F, S\}$.

The cost index can be interpreted as follows. Suppose the central planner spends a budget of m_j (per agent) to reimburse the cost of service j. Denote by $U_j(m_j)$ the total utility loss from service j. We can derive

$$\frac{\partial U_j(m_j)}{\partial m_j} = -\frac{\mathsf{E}[u'((1-r_j)C_j)C_j]}{\mathsf{E}[C_j]}, \tag{12}$$

where r_j is the reimbursement ratio determined by the budget constraint. The absolute value $|\partial U_j(m_j)/\partial m_j|$ represents the marginal benefit of increasing budget for service j. By the convexity of $u(l)$, the right-hand side of (12) decreases in the current reimbursement ratio r_j. Thus, the cost index b_j represents the maximum marginal benefit from investing in service j, which is obtained when we start to reimburse the service ($r_j = 0$).

When the central planner sets the reimbursement ratios r_F and r_S, she is actually allocating the budget between the two services. If both services are reimbursed at the optimal policy ($r_F, r_S > 0$), they must have the same marginal benefit $|\partial U_j(m_j)/\partial m_j|$. Otherwise, the central planner can improve the objective by transferring the budget from the service with smaller marginal benefit to the larger one.[2] This condition determines the optimal ratio reimbursement policy, which is summarized in the following proposition.

Proposition 2 *The optimal ratio reimbursement policy (r_F^*, r_S^*) has the following structure. There exists a threshold m_r such that for budget $m \in [0, m_r]$, only the service $j \in \{F, S\}$ with higher cost index in (11) is reimbursed. For $m \in [m_r, \bar{m}]$, both services F and S are reimbursed. Their reimbursement ratios satisfy*

$$\frac{\mathsf{E}[u'((1-r_F^*)C_F)C_F]}{\mathsf{E}[C_F]} = \frac{\mathsf{E}[u'((1-r_S^*)C_S)C_S]}{\mathsf{E}[C_S]}. \tag{13}$$

Both the reimbursement ratios r_F^* and r_S^* are non-decreasing in the budget m. The threshold m_r is explicitly characterized in Section EC.3.3.

Proposition 2 reveals the structure of the optimal ratio reimbursement policy (r_F^*, r_S^*). When the budget level is relatively low ($m < m_r$), the optimal policy is to use all budget to reimburse the the service with higher cost index, which is associated with higher or more dispersed cost. The other service is unreimbursed. When the budget level exceeds m_r, both services get reimbursed by the central planner. Their reimbursement ratios r_F^* and r_S^* increase in the total budget, suggesting no agent is worse off under a higher budget level. Such structure is in line with practice. For example, with limited funding, the inpatient service is usually reimbursed before the outpatient service.

This structure is economically intuitive. Suppose $b_S > b_F$, then the central planner would first reimburse the second-level service as it leads to higher marginal benefit. As the budget level increases, the reimbursement ratio r_S increases. By (12), this decreases the marginal benefit of increasing the budget for the

[2] The assumption $u'(0) = 0$ ensures that we will not fully reimburse one service while partially reimburse the other one at the optimal policy. See proof in Section EC.3.3.

second-level service. When $m > m_r$, the marginal benefit of investing in the second-level service becomes low enough such that it is beneficial to reimburse both types of services. In this regime, the marginal benefit of allocating budget to the two services equals each other as in (13). Thus, the central planner cannot improve the objective by changing the allocation of budget.

Combining the optimality condition with the budget constraint allow us to solve the optimal (r_F^*, r_S^*). As shown in Section EC.3.3, the results can be further simplified under the power utility loss function in (3). In this case, the optimal r_F^* and r_S^* can be solved explicitly, and both increase piecewise-linearly in the budget m. An example of the optimal (r_F^*, r_S^*) is illustrated in the right panel of Figure 1. We see that both services are reimbursed when the budget level is high enough.

Denote the total utility loss under the optimal cap and ratio reimbursement policies by $U_c(m)$ and $U_r(m)$, respectively. We have the following proposition.

Proposition 3 *Under both cap and ratio reimbursement policies, the single-period total utility loss $U_c(m)$ and $U_r(m)$ are convex and decreases in budget level m. In addition, the utility loss is larger under the ratio policy: $U_r(m) \geq U_c(m)$ for all m.*

From the central planner's aspect, the optimal cap reimbursement policy always leads to better system performance, i.e., smaller total utility loss, than the ratio policy.[3] This is not surprising as the cap policy is the optimal solution to the unconstrained optimization problem (4), while the ratio policy additionally imposes the linear functional form (10). Intuitively, the cap policy is more effective in avoiding extreme costs faced by agents. However, it is not always preferred by individual agents. For agents with relatively low cost, the ratio policy may deliver a larger reimbursement amount (thus a smaller net cost). In contrast, the agents with high costs benefit more from the cap policy.

Under both the optimal cap and ratio policies, the total utility loss is convex decreasing in the budget level m. Thus, the marginal benefit of increasing the budget drops as the budget level increases. This can be explained by the convexity of the utility loss function $u(l)$, which reflects the risk aversion of the central planner. With a larger budget, the agent's net cost decreases and the impact on her utility loss becomes smaller.

The cap and ratio policies are relatively easy to implement. Thus, the two policies, as well as their different hybrids, are widely used in practice. For example, Hong Kong and U.K. only charge small fixed costs for the outpatient and inpatient service in their public health systems. Canada has a free healthcare system for basic medical services. These can be viewed as examples of the cap reimbursement policy. The ratio reimbursement policy is widely used in France and mainland China, in which the reimbursement ratios depend on the service types and insurance schemes. Many places adopt hybrid reimbursement models that

[3] When the costs are constants, the optimal cap and ratio policies coincide. This is because with only two possible cost levels, any reimbursement policy can be represented by the corresponding ratios (r_F, r_S).

combine cap and ratio reimbursement policies. In Japan, South Korea, and Spain, all costs above a cumulative threshold as well as a proportion of costs below the threshold are reimbursed. In Taiwan, outpatients only pay a fixed fee, while inpatients need to pay a certain percentage of their medical cost, depending on the length of stay.[4]

Both the cap and ratio policies have their pros and cons. The cap policy leads to smaller utility loss as it benefits agents with relatively high costs. However, it leads to potential moral hazard — agents may require unnecessary medical services given their self-payment is capped (e.g., Pauly 1968, Ata et al. 2012). This is mitigated by the ratio policy, as the net amount spent by an agent always increases in her raw service cost. In this work, we use the cap and ratio policies as two fundamental and representative settings. More complex reimbursement policy designs (e.g., mixing the cap and ratio policies) and responses in agents' behaviors are deferred to future research.

3. Multi-Groups Model with Horizontal Pooling

In this section, we extend our model to the multi-group setting, i.e., the central planner manages the medical reimbursement policy of multiple groups. We consider different pooling systems, depending on whether the budgets are shared and/or a common reimbursement policy is used.

3.1. Model Set-up and Pooling Systems

In the model, there are K groups of agents indexed by $i = 1, 2, ..., K$. There are still two types of services F and S. For group i, its population mass, per capita budget level, and services incidence in (6) are denoted by $N^{(i)}$, $m^{(i)}$, and $(h_F^{(i)}, h_S^{(i)})$, respectively. Let the services costs for group i be $C_F^{(i)}$ and $C_S^{(i)}$, with distributions $G_F^{(i)}$ and $G_S^{(i)}$. This captures multiple types of heterogeneity across the groups: they can have different population sizes, budget levels, services incidences, and cost distributions once the service is required. For example, one group may be on average healthier than the other (young versus senior), thus having lower services incidences and costs. Or, one group may enjoy better benefit package and financial protection from the insurance, which is reflected by a higher per capita budget level.

Managing the medical insurance for multiple groups is a significant challenge for many healthcare authorities. Providing universal and equitable healthcare access and service for entire population has been a fundamental goal for most modern governments (Culyer and Newhouse 2000). However, in reality, medical insurance systems are often fragmented and operated separately with different expenditure levels, benefit

[4]Some related reimbursement policies can be found in the following links. HK: https://www.ha.org.hk/visitor/ha_visitor_index.asp?Content_ID=10045&Lang=ENG; UK: https://www.nhs.uk/nhs-services/hospitals/about-nhs-hospital-services/; German: https://howtogermany.com/insurance/health-insurance/paying-medical-expenses-health-insurance-claims-germany/; France: https://www.ameli.fr/assure/remboursements/rembourse/tableau-recapitulatif-taux-remboursement; Taiwan: https://law.moj.gov.tw/LawClass/LawAll.aspx?pcode=L0060001; Japan: https://www.med.jrc.or.jp/en/tabid/835/Default.aspx.

packages, and reimbursement policies. This is especially true for developing countries. Such fragmentation greatly hinders the healthcare access and financial protection of the susceptible groups and leads to welfare loss for the society. To overcome this challenge, integration and consolidation of medical insurance schemes have become a widely adopted goal of healthcare policy makers, and practical actions have been taken (e.g. Atun et al. 2015, Cotlear et al. 2015). For example, in mainland China, the basic medical insurance schemes for rural and urban residents has been gradually consolidated since 2016 (Ma 2021).

In the context of our model, we study the effects of pooling in the reimbursement problem. Pooling is one of the most important strategies in operations management. It is widely used in different areas such as revenue management (Afeche and Pavlin 2016), supply chain (Qi et al. 2015), inventory management (Corbett and Rajaram 2006), and healthcare operations (Song et al. 2020). We consider three systems of horizontal pooling in the multigroup setting. In the *non-pooling* system (NP), each group operates separately with its own budget and reimbursement policy. In the *full pooling* system (FP), the central planner pools the budget from all groups, and use a common reimbursement policy for all of them. In the *monetary pooling* system (MP), the central planner still pools the budgets together, but is allowed to use heterogeneous reimbursement policies for the groups. The settings of the three systems are summarized in Table 1.

Table 1 Settings of Three Horizontal Pooling Systems

	Budget sharing	Policy flexibility	Decision variable	Budget constraint
Non-pooling	No	Yes	$\{\phi_S^{(i)}, \phi_F^{(i)}\}_{i=1}^K$	(15)
Full pooling	Yes	No	ϕ_S, ϕ_F	(16)
Monetary pooling	Yes	Yes	$\{\phi_S^{(i)}, \phi_F^{(i)}\}_{i=1}^K$	(16)

The three pooling systems reflect two practically important aspects in the multi-group setting: resource sharing and policy flexibility. In non-pooling, each group operates independently without coordination from the central planner. This mirrors the decentralized systems in revenue management (e.g., Netessine and Rudi 2006, Yang et al. 2021). In full pooling, all groups are integrated into one with a shared budget and a common reimbursement policy. Thus, it allows resource sharing but loses the flexibility in policy design. For example, we cannot set higher reimbursement ratios for groups that have higher incidence and medical costs. Monetary pooling achieves both resource sharing and flexibility in reimbursement policies. This leads to the best system performance in the three systems. However, it may be more challenging to implement in reality.

The central planner minimizes the total utility loss of all groups, which can be expressed as

$$U(\mathbf{m}) = \min \sum_{i=1}^{K} \sum_{j\in\{F,S\}} w^{(i)} h_j^{(i)} \mathsf{E}\left[u\left(l_j^{(i)}(C_j^{(i)})\right)\right], \tag{14}$$

where $w^{(i)}$ represents the population weight of group i, defined as $w^{(i)} = N^{(i)}/\sum_{j=1}^{K} N^{(j)}$; the net cost function $l_j^{(i)}(x) := x - \phi_j^{(i)}(x)$ for each group i and service type j. In the non-pooling and monetary pooling systems, the decision variables are $(\phi_S^{(i)}, \phi_F^{(i)})$ for $i = 1, 2, ..., K$, as different reimbursement policy can be used in each group. In the full pooling system, the central planner only decides the common reimbursement policy (ϕ_S, ϕ_F) for all groups. We consider admissible reimbursement functions in (1) all cases.

In the non-pooling system, the budget constraint is imposed on each group:

$$\sum_{j\in\{F,S\}} h_j^{(i)} \mathsf{E}\left[\phi_j^{(i)}\left(C_j^{(i)}\right)\right] = m^{(i)}, \quad \forall i = 1, 2, ..., K. \tag{15}$$

In the full pooling and monetary pooling systems, the central planner only needs to satisfy the constraint based on the total expense and budget of all groups, which is

$$\sum_{i=1}^{K} \sum_{j\in\{F,S\}} w^{(i)} h_j^{(i)} \mathsf{E}\left[\phi_j\left(C_j^{(i)}\right)\right] = \sum_{i=1}^{K} w^{(i)} m^{(i)}. \tag{16}$$

The last two columns of Table 1 summarize the decision variables and budget constraints in the three pooling systems. The explicit formulations of the optimization problems are provided in Section EC.1.

3.2. Optimal Policy in Pooling Systems

We solve the optimal reimbursement policy in the pooling systems. As in Section 2, we consider both the unconstrained optimal policy and the optimal ratio policy. In the non-pooling system, the optimal reimbursement policies can be solved for the K groups separately as in Propositions 1 and 2. In the full-pooling system, the central planner essentially combines all groups into one. Thus, we can apply our single-group results with the new parameters and cost distributions based on the aggregation of the K groups. To save space, we detail the optimal solution under full pooling in Section EC.2.1.

In the follows, we look into the optimal reimbursement policy in the monetary pooling system. Under monetary pooling, different reimbursement policy can be used for each group, although the budgets are pooled. We first consider the unconstrained optimal policy without imposing a specific form of the reimbursement function $\phi_j^{(i)}(x)$. Next, we solve the optimal ratio reimbursement policy with $\phi_j^{(i)}(x) = r_j^{(i)} x$. Here we only need to decide the reimbursement ratios $(r_F^{(i),*}, r_S^{(i),*})_{i=1}^K$ for the two services of each group. The results are summarized in the following proposition.

Proposition 4 *(i) Under monetary pooling, the optimal reimbursement policy has a cap structure, with the reimbursement function given by*

$$\phi_j^{(i),*}(x) = \max\{x - \tau^*, 0\}, \quad \textit{for } j \in \{F, S\} \ \textit{ and } \ i = 1, 2, ..., K. \tag{17}$$

The self-payment threshold τ^* convex decreases in the total budget $\sum_{i=1}^{K} w^{(i)} m^{(i)}$.

(ii) For the optimal ratio policy $(r_F^{(i),*}, r_S^{(i),*})_{i=1}^K$ *under monetary pooling, there exists a threshold* $\tilde{b}$ *such that*

$$r_j^{(i),*} > 0 \quad \textit{if and only if} \quad b_j^{(i)} > \tilde{b},$$

where $b_j^{(i)}$ *is the cost index in (11) for service* $j \in \{F, S\}$ *in Group* i*. For all reimbursed services with* $r_j^{(i),*} > 0$*, their reimbursement ratios satisfy*

$$\frac{\mathsf{E}[u'((1 - r_j^{(i),*})C_j^{(i)})C_j^{(i)}]}{\mathsf{E}[C_j^{(i)}]} = \tilde{b}. \tag{18}$$

The reimbursement ratios are non-decreasing in the total budget $\sum_{i=1}^K w^{(i)} m^{(i)}$.

Proposition 4 shows that when there is no restriction on the form of the reimbursement policy, the central planner should still use a "max-out-of-pocket" (cap) policy under the monetary pooling. The threshold τ^* for self-payment is same for all service types and groups. The intuition is straightforward. Given the convexity of the utility loss function, the central planner should reimburse the agents in the descending order of their actual costs, regardless of their group and service type. As in the single-group setting, the threshold τ^* convex decreases in the total budget level.

Under monetary pooling, the structure of the optimal ratio policy is similar to the single-group results. For service type j of group i, the central planner does not reimburse it if its cost index $b_j^{(i)}$ is smaller than the threshold $\tilde{b}$. Otherwise, its reimbursement ratio $r_j^{(i),*}$ is positive and satisfies (18). Economically, with the optimal ratio policy, the central planner only reimburse the services with high enough marginal benefit, i.e., above the the threshold $\tilde{b}$. In addition, all reimbursed services must have the same marginal benefit from increasing their own budget. Thus, the central planner cannot further improve the objective by changing the allocation across groups and services. As the total budget increases, more services get reimbursed and the reimbursement ratios increase.

Corollary 1 *Assume the power utility loss function (3). Under monetary pooling, the optimal reimbursement ratios* $(r_F^{(i),*}, r_S^{(i),*})_{i=1}^K$ *are given by*

$$r_j^{(i),*} = \max\left\{1 - (\tilde{b})^{\frac{1}{(\gamma-1)}} / (b_j^{(i)})^{\frac{1}{(\gamma-1)}}, 0\right\}. \tag{19}$$

In addition, $(r_F^{(i),*}, r_S^{(i),*})_{i=1}^K$ *increase piecewise linearly in the total budget* $\sum_{i=1}^K w^{(i)} m^{(i)}$*, with break points occurring when a new service starts to get reimbursed.*

With the power utility loss function (3), the optimal ratio policy under monetary pooling can be further simplified. By (19), the optimal reimbursement ratios $r_j^{(i),*}$ are higher for services with larger cost indexes $b_j^{(i)}$. In addition, they increase piecewise linearly in the total budget, same as the pattern in the right panel of

Figure 1. In this case, the optimal ratios $(r_F^{(i),*}, r_S^{(i),*})_{i=1}^K$ can be solved in closed-form based on the ranking of the services' cost indexes. The explicit expressions are provided in Section EC.4.2.

3.3. Performance of Different Pooling Systems

In this section, we compare the system performance under the three pooling systems. Understanding the benefits and limitations of different pooling systems is practically important for the central planner in designing the reimbursement policy for multiple groups.

Denote the total utility loss in (14) under non-pooling, full pooling, and monetary pooling by $U^{(NP)}$, $U^{(FP)}$, and $U^{(MP)}$, respectively. Clearly, pooling homogeneous groups together has no impact on the system performance, as all groups use the same optimal reimbursement policy before and after pooling. This is shown in the following remark.

Remark 2 *If the K groups are fully homogeneous with same services incidence, cost distributions and budget level. Then horizontal pooling does not lead to any benefit compared with non-pooling, i.e., $U^{(NP)} = U^{(FP)} = U^{(MP)}$. This holds for both cap and ratio reimbursement policies.*

The analysis becomes more complex when pooling heterogeneous groups. It is easy to see that monetary pooling always leads to non-worse outcomes than full pooling and non-pooling, as the optimal policies in the latter two systems are always feasible in monetary pooling. However, two important questions still remain. The first is when it is optimal to use a common reimbursement policy for all groups. In this case, policy flexibility brings no value, and full pooling achieves the same performance as monetary pooling. The second is how should the central planner choose between policy flexibility and resource sharing. This relies on the performance comparison between full pooling and non-pooling, i.e., whether it is better to fully integrate the groups into one or manage them separately.

In Theorem 1, we provide analytical answers to the two questions regarding the performance of the pooling systems. We first introduce the following definition for the service cost distributions across groups.

Definition 2 *The service costs are homogeneous if the cost distributions are the same across all groups for both types of services. Otherwise, the service costs are heterogeneous across groups, i.e., there exists $j \in \{F, S\}$, i_1, $i_2 \in \{1, 2, ..., K\}$ such that the distributions of $C_j^{(i_1)}$ and $C_j^{(i_2)}$ are different.*

Theorem 1 *(i) The optimal cap reimbursement policy under monetary pooling is same for all groups $i \in \{1, 2, ..., K\}$. The total utility loss satisfies: $U_c^{(NP)} \geq U_c^{(FP)} = U_c^{(MP)}$.*

(ii-a) With homogeneous service costs, the optimal ratio reimbursement policy $(r_F^{(i),*}, r_S^{(i),*})$ under monetary pooling is same for all groups $i \in \{1, 2, ..., K\}$, i.e.,

$$r_F^{(i),*} = r_F^*, \quad r_S^{(i),*} = r_S^*, \quad \forall i = 1, 2, ..., K,$$

where (r_F^*, r_S^*) *is the optimal ratio policy under full pooling. The total utility loss satisfies:* $U_r^{(NP)} \geq U_r^{(FP)} = U_r^{(MP)}$.

(ii-b) With heterogeneous service costs, the optimal ratio reimbursement policy $(r_F^{(i),*}, r_S^{(i),*})$ *under monetary pooling may differ across the groups. The total utility loss satisfies* $U_r^{(NP)} \geq U_r^{(MP)}$ *and* $U_r^{(FP)} \geq U_r^{(MP)}$*;* $U_r^{(FP)}$ *may be larger or smaller than* $U_r^{(NP)}$.

Table 2 Utility Loss under Different Pooling Systems

	Homogeneous Cost	Heterogeneous Cost
Cap Policy	$MP = FP \succeq NP$	$MP = FP \succeq NP$
Ratio Policy	$MP = FP \succeq NP$	$MP \succeq FP, NP$ and $NP \lesseqgtr FP$

For ease of visualization, the main results in the theorem are also summarized in Table 2, in which "$=$" means the same performance; "$\succeq$" means the left-hand side outperforms the right-hand side with a smaller total utility loss; "$\lesseqgtr$" means indefinite relation.

Theorem 1 establishes sufficient conditions for full pooling to achieve the same performance as monetary pooling. In these cases, policy flexibility is not needed and a common reimbursement policy is optimal. We have the following conditions. First, with the optimal cap policy, full pooling always has the same performance as monetary pooling regardless of the cost distributions. This is because the central planner always sets a universal self-payment threshold τ^* for all groups and services in the optimal cap policy (see Proposition 4). Second, with the optimal ratio policy, full pooling and monetary pooling have the same performance when all groups have homogeneous cost distributions for both service types. Economically, Proposition 4 shows that once the total budget is fixed, the optimal reimbursement ratios are solely determined by the service cost distributions. Thus, the group-specific policy is not needed with homogeneous service costs.

The above results have practical implications for policy makers. It shows that if the cap policy is used, or the ratio policy is used with homogeneous costs, the central planner does not need to implement different policies for the groups. Instead, it is sufficient to consolidate the groups as one and apply a common reimbursement policy. This holds even the groups are very different in their population sizes, budget level, and service incidences. A common reimbursement policy is more convenient to implement by the central planner. In addition, it achieves horizontal equity and fairness across the groups (Culyer and Newhouse 2000). We show that this can be done without hurting the system performance under relatively general conditions.

Theorem 1 also sheds lights on the comparison between non-pooling and full pooling, i.e., the trade-off between policy flexibility and resource sharing. We find that with homogeneous service costs or cap reimbursement policy, full pooling always outperforms non pooling regardless of the population sizes, services incidence, and budget levels of the groups. In these situations, the central planner should simply integrate all groups into one rather than manage them separately. However, when ratio reimbursement policy is used for

groups with heterogeneous services costs, it is unclear whether non-pooling or full pooling leads to better system performance. In this case, the benefit of using flexible policy may outweigh that of budgets sharing. For example, it may be more desirable for the central planner to set higher reimbursement ratios for the group with higher average costs, even only with its own separate budget.

In the following proposition, we use a simplified example to show that both non pooling and full pooling can strictly outperform the other. We consider two groups with same population size N and service incidences (h_F, h_S). We assume the per capita budget is given by m for Group 1 and $k_m m$ for Group 2 with $k_m \geq 1$. The service costs are constants (c_F, c_S) for Group 1 and $(k_c c_F, k_c c_S)$ for Group 2, i.e., the service cost vectors are parallel for the two groups. The parameters k_m and k_c measures the heterogeneity in the two groups' budget and cost, respectively. For example, a larger value of k_c may be because the agents in Group 2 are less healthy than those in Group 1, therefore having higher medical costs.

Proposition 5 *Assume power utility loss function (3) and optimal ratio reimbursement policy. We have the following comparison between non-pooling and full pooling:*

(i) If $m \leq h_S(c_S - c_F)$ and $k_m m \leq k_c h_S(c_S - c_F)$, there exists $k_1 \in (1, k_m)$ such that full pooling outperforms non-pooling for

$$k_c \in (0, k_1) \ \text{ or } \ k_c \in (k_m, \infty);$$

and non-pooling outperforms full pooling for $k_c \in (k_1, k_m)$.

(ii) If $m \geq h_S(c_S - c_F)$ and $k_m m \geq k_c h_S(c_S - c_F)$, there exists $k_2 \in (1, k_m)$ such that full pooling outperforms non-pooling for

$$k_c \in (0, k_2) \ \text{ or } \ k_c \in (k_m, \infty);$$

and non-pooling outperforms full pooling for $k_c \in (k_2, k_m)$. The explicit expressions of k_1 and k_2 are given in Section EC.4.4.

The first (resp. second) case in Proposition 5 corresponds to the scenario where only service S (resp. both services) is reimbursed in both groups in the non-pooling system. We see that full pooling strictly outperforms non pooling when k_c is relatively small and large, but underperforms with a medium k_c. The results can be interpreted as follows. First, when we pool the two groups with $k_c > k_m$, it is easy to verify that the fund flows from Group 1 to Group 2, which has higher marginal benefits from increasing the budget due to its higher costs. This improves the overall system performance. On the other hand, when $k_c < k_1$ in (i) or $k_c < k_2$ in (ii), the service costs of Group 2 are relatively low relative to its budget level. In this case, budget sharing reduces the imbalance between the two groups and substantially improves the performance of Group 1. Intuitively, full pooling outperforms non pooling when the high-cost group is more budget constrained or when the cost-budget imbalance is significant across groups.

Our results suggest that a key factor in designing the reimbursement policy is the distribution of service costs in each group. This should be differentiated from the service incidence, which measures the probability of requiring the service. If a cap policy is used, the cost heterogeneity does not hurt in full pooling. This is because a cap policy always reimburses the agents using a fixed self-payment threshold, regardless of their actual costs. However, if the ratio policy is used, cost heterogeneity brings value to the policy flexibility when managing multiple groups. In reality, heterogeneous cost distributions can occur when the groups have different severity levels and care demands for given medical conditions. For example, the agents in one group may include more senior citizens with worse health conditions. If we define groups by regions, then the geographical variation in disease may also lead to heterogeneous service costs. In such cases, it would be valuable for the central planner to set different ratio reimbursement policies for each group.

4. Dynamic Model with Intertemporal Pooling

In this section, we develop a dynamic model in which the central planner needs to make reimbursement decisions in multiple periods. In the dynamic setting, the central planner cares about the utility loss in both current and future periods. Thus, she needs to balance between spending more for current period versus saving more for the future. This introduces an intertemporal trade-off in the decision-making. In reality, central planner faces such trade-off when she manages a medical fund to reimburse the services over time. We can view the dynamic model as a form of intertemporal pooling, in the sense that the fund is shared across multiple periods.

4.1. Dynamic Model Set-up

We model the reimbursement problem in the dynamic setting by a general Markov decision process (MDP). We consider an infinite time horizon discretized into equal-length periods. The central planner manages an insurance fund that is used exclusively for reimbursing the agent's medical costs in current and future periods. Similar to the set-up in Section 2, there is a single group of agents with two types of service: first-tier (F) and second-tier (S) medical services, with the latter being on average more costly than the first one. The service incidence are h_F and h_S. The two types of services incur costs of C_F and C_S without reimbursement, which are random variables with finite supports. We assume the population size, service incidences, and the cost distributions do not vary with time.

State, action, and transition: We index the periods by $t = 0, 1, 2, ...$ At the beginning of each period t, the system state is represented by the per capita fund level s_t, which is the amount of available wealth in the fund. We require the fund level s_t to be non-negative for all time, i.e., the fund cannot go bankruptcy. Thus, the state space is given by $\mathcal{S} := [0, \infty)$.

After observing the state s_t, the central planner makes the following decisions. First, she decides the spending amount $m_t \in [0, s_t]$ for reimbursing the medical service costs in current period. Then, with the

budget m_t, the central planner decides the reimbursement policy (ϕ_F, ϕ_S) as described in Section 2. That is, if a patient incurs cost x from medical service $j \in \{F, S\}$, the central planner reimburses the amount of $\phi_j(x)$. Thus, the central planner's action can be represented by a tuple $a = (m, \phi_F(\cdot), \phi_S(\cdot))$. For state s, the admissible action space is given by

$$\mathcal{A}(s) = \{(m, \phi_F(\cdot), \phi_S(\cdot)) : 0 \le m \le s \text{ and } (\phi_F, \phi_S) \in \Phi(m)\}. \tag{20}$$

The constraint $m \in [0, s]$ states that the spending amount for current period's reimbursement cannot exceed the total fund level. The feasible set $\Phi(m)$ for reimbursement policy is defined as

$$\Phi(m) = \{(\phi_S, \phi_F) : \sum_{j \in \{F,S\}} h_j \mathsf{E}[\phi_j(C_j)] = m \text{ and } (\phi_F, \phi_S) \in \mathcal{C}\}, \tag{21}$$

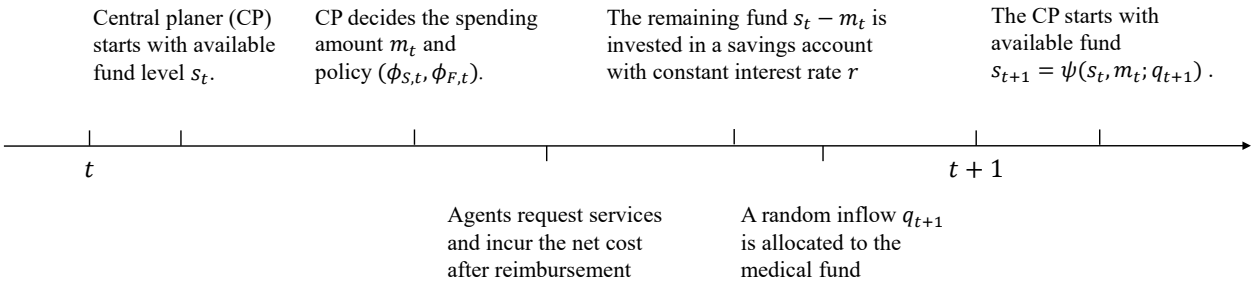
where h_F and h_S are the services incidence. As discussed in Section 2, it requires that the total reimbursement expense must equal the budget level m, and the reimbursement amount must be non-negative and no larger than the actual expense.

After taking the action $a_t = (m_t, \phi_{F,t}, \phi_{S,t})$, the system evolves as follows. First, the remaining fund with amount $s_t - m_t$ is put into a savings account and earns an interest rate of r. Second, a random new per capita inflow q_{t+1} is allocated to the medical fund. Thus, the fund level at the beginning of next period is given by

$$s_{t+1} = \psi(s_t, m_t; q_{t+1}) := (1 + r)(s_t - m_t) + q_{t+1}, \tag{22}$$

where $\psi(s, m; q)$ is the transition function. We assume q_t is independent and identically distributed across periods. In addition, it takes value in a finite non-negative interval $[\underline{q}, \overline{q}]$ with continuous probability density function $f_q(\cdot)$. The timing of decision and system transition are summarized in Figure 2.

Figure 2 Timeline for Dynamic Model



Cost and objective. With state s and action a, the total utility loss in current period is given by

$$\tilde{u}(s, a) := h_F \mathsf{E}[u(l_F(C_F; a))] + h_S \mathsf{E}[u(l_S(C_S; a))] \text{ for } a = (m, \phi_F, \phi_S) \in \mathcal{A}(s), \tag{23}$$

where $l_j(x;a) := x - \phi_j(x;a)$ for $j \in \{F, S\}$ denotes the unreimbursed medical cost under action a. The expectations in (23) are taken over the random costs C_F and C_S. The central planner minimizes the total discounted utility loss over an infinite horizon by choosing a policy π that maps the state space to admissible action, i.e., $\pi : \mathcal{S} \to \mathcal{A}$. The value function can be represented as:

$$v(s) := \min_{\pi} \mathsf{E}_{\mathbf{q}} \left[\sum_{t=0}^{\infty} \beta^t \, \tilde{u}(s_t, \pi(s_t)) \middle| \, s_0 = s \right], \tag{24}$$

where $\beta \in (0,1)$ is the discount factor. The expectation $\mathsf{E}_{\mathbf{q}}$ accounts for the random realizations of inflow $\mathbf{q} = (q_1, q_2, ...)$. Given the sequence $\mathbf{q}$ and the policy π, the transition of states $\{s_t\}$ can be fully determined. Denote the optimal policy by $\pi^* = (m^*(s), \phi_F^*(\cdot; s), \phi_S^*(\cdot; s))$, which minimizes the objective. For the infinite horizon problem (24), it suffices to consider the stationary policies (Puterman 1994).

To solve the optimal policy π^* in the dynamic model, we notice that the per-period utility loss $\tilde{u}(s,a)$ in (23) is independent of the state s. On the other hand, the transition of state in (22) only depends on the spending amount m_t, but not the reimbursement policy $\phi_{F,t}(\cdot)$ and $\phi_{S,t}(\cdot)$. This suggests that the central planner's decision can be decomposed to two stages. She first decides the current period's spending amount m_t for reimbursing the medical costs. This simultaneously determines the savings amount $s_t - m_t$ for future periods. Then, the central planner decides the optimal reimbursement policy $\phi_{F,t}$ and $\phi_{S,t}$ given the budget level m_t. In the second stage, the central planner essentially faces a single-period optimization problem. The following lemma formalizes such intuition.

Lemma 2 *The optimal policy $\pi^*(s) = (m^*(s), \phi_F^*(\cdot; s), \phi_S^*(\cdot; s))$ has the following property. Given the budget level $m^*(s)$, the policy $(\phi_F^*(\cdot; s), \phi_S^*(\cdot; s))$ is the solution to the single-period utility loss minimization problem:*

$$U(m^*(s)) = \min_{(\phi_F, \phi_S) \in \Phi(m^*(s))} h_F \mathsf{E}\big[u(l_F(C_F))\big] + h_S \mathsf{E}\big[u(l_S(C_S))\big] \tag{25}$$

with $l_j(x) = x - \phi_j(x)$. That is, the optimal reimbursement policy depends on state s only via the budget level $m^*(s)$.

Given the budget level $m^*(s)$, the single-period optimization problem (25) has been explicitly solved in Propositions 1 and 2 for the cap and ratio reimbursement policies, respectively. Thus, it suffices to focus on the optimal spending amount $m^*(s)$ for each state s in the dynamic problem. In light of this, the value function in (24) can be rewritten as:

$$v(s) = \min_{\pi_m} \mathsf{E}_{\mathbf{q}} \left[\sum_{t=0}^{\infty} \beta^t U(\pi_m(s_t)) \middle| \, s_0 = s \right], \tag{26}$$

where the new policy function π_m maps the state s to the spending amount $m \in [0, s]$. For the dynamic programming problem (26), we can derive its Bellman's equation as

$$v(s) = \min_{m \in [0,s]} \left[U(m) + \beta \mathsf{E}_q v(\psi(s, m; q)) \right]. \tag{27}$$

Thus, the optimal spending amount $m^*(s)$ satisfies

$$m^*(s) \in \arg\min_{m \in [0,s]} \left[U(m) + \beta \mathsf{E}_q v(\psi(s, m; q)) \right]. \tag{28}$$

Note that in the optimal policy, we never need to invest more than the maximum amount $\bar{m}$ in (7), with which all costs can be fully reimbursed.

4.2. Optimal Policy in Dynamic Model

In this section, we investigate the structure of optimal policy in our dynamic model. By Proposition 3, the per-period utility loss $U(m)$ is convex and decreases in the budget level m. Based on this, we can develop the following results for the value function $v(s)$.

Proposition 6 *Under both cap and ratio reimbursement policies, the value function $v(s)$ convex decreases in the medical fund level s.*

Proposition 6 states that the total discounted utility loss over the infinite horizon is smaller when the central planner starts with more wealth in the medical fund. This reveals the expected intertemporal trade-off in the central planner's decision making — she needs to balance between spending more for reimbursing current period's medical costs versus saving more for future periods. In addition, the convexity of $v(s)$ implies that the marginal benefit of increasing s becomes smaller as the fund level increases. Such a pattern essentially stems from the convexity of the per-period utility loss $U(m)$, as established in Proposition 3.

We solve the optimal policy $m^*(s)$ for a given state s. When $\beta(1+r) = 1$ and inflow q is constant, the optimal spending amount $m^*(s)$ can be solved in closed-form by the choice model in economics theory (Sachs and Larrain B. 1993). In this special case, the optimal budget level $m^*(s_t)$ and reimbursement policy $(\phi^*_{F,t}, \phi^*_{S,t})$ are stationary over time. This is summarized in the following lemma.

Lemma 3 *If $\beta(1+r) = 1$ and inflow q is constant in each period, the optimal spending amount $m^*(s)$ can be solved as:*

$$m^*(s) = \min \left\{ s, \frac{r}{1+r} s + \frac{1}{1+r} q \right\}.$$

This applies to both cap and ratio reimbursement policies. In this case, the fund level s_t is stationary with $s_t \equiv \max\{s_0, q\}$ for $t \geq 1$.

In general cases, closed-form solution of $m^*(s)$ is unavailable due to the analytical complexity embedded in the dynamic setting. Thus, we focus on characterizing the structure of the optimal policy. We first develop the following theorem on how the optimal spending amount $m^*(s)$ changes with the fund level s.

Theorem 2 *For both cap and ratio reimbursement policies, the optimal spending amount $m^*(s)$ is continuous and increases in the medical fund level s. There exists a threshold $\tilde{s}_l \geq 0$ such that the optimal spending amount $m^*(s)$ satisfies:*

(i) if $s \leq \tilde{s}_l$, we have $m^*(s) = 0$ or $m^*(s) = s$. That is, the central planner should either fully spend for current period reimbursement or fully save for the future. In addition, $m^*(s) = s$ holds if $\beta(1+r) \leq 1$.

(ii) if $s > \tilde{s}_l$, we have $0 < m^*(s) < s$. That is, the central planner should partially spend and save. In addition, under the assumption that $\mathsf{Pr}(q < \bar{m}) > 0$, the optimal spending amount satisfies $m^*(s) < \bar{m}$ at $s = \bar{m}$.

Theorem 2 shows that when the current fund level s is relatively low, the central planner would either spend all the wealth for current period's reimbursement or save all for the future. Furthermore, if discount factor is large or interest rate is low ($\beta(1+r) \leq 1$), the optimal action is to spend all with $m^*(s) = s$. As the fund level s increases, the optimal spending amount $m^*(s)$ also increases. When $s > \tilde{s}_l$, the central planner would spend a part of the fund for current period and save the rest for the future, reflecting the intertemporal trade-off in the decision making. Under the natural assumption that per period inflow may be insufficient to fully cover the medical expense ($\mathsf{Pr}(q < \bar{m}) > 0$), the central planner will not spend all fund for current period at $s = \bar{m}$, although this is admissible. The structure in Theorem 2 is indeed observed in Lemma 3 under the special case of $\beta(1+r) = 1$ and constant inflow.

In the follows, we further consider a special case of the utility loss function — the power function with $\gamma = 2$ in (3). In this case, we can obtain more explicit analytical property of the optimal spending amount $m^*(s)$. The results provide valuable managerial insights on designing the reimbursement policy in the dynamic setting. Without loss of generality, we assume the cost indexes satisfy $b_F < b_S$. Recall the budget level m_r defined in Proposition 2: If the budget $m < m_r$, only service S is reimbursed under the optimal ratio policy. If $m > m_r$, the central planner reimburses both services.

Theorem 3 *Under power utility loss function (3) with $\gamma = 2$, the optimal spending amount $m^*(s)$ has the following structure with respect to the fund level s. Define $\underline{s} := \sup\{s : m^*(s) = 0\}$.*

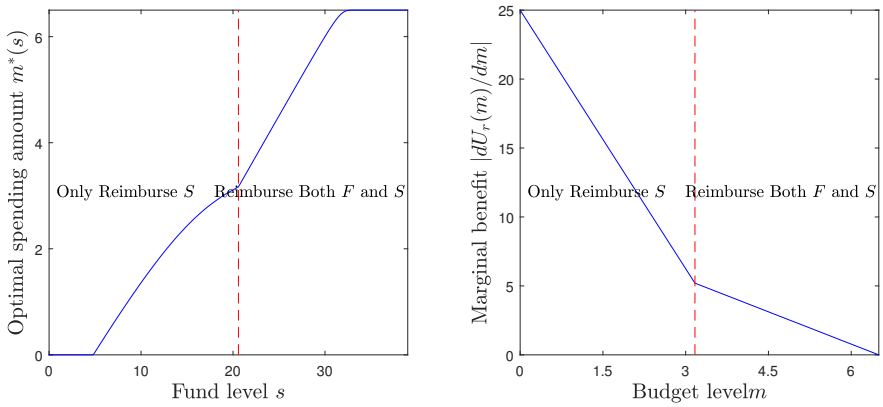
(i) Under ratio reimbursement policy, define the state s_r by $m^*(s_r) = m_r$. The optimal spending amount $m^*(s)$ is concave increasing in the regions $(\underline{s}, s_r)$ and (s_r, ∞). At the break point s_r, the one-side first-order derivative satisfies $(m^*)'_+(s_r) \leq (m^*)'_-(s_r)$.

(ii) Assume that the service costs C_S and C_F take random discrete values $c_{(1)} < c_{(2)} < \cdots < c_{(n)}$. Under cap reimbursement policy, define the states $s_{(i)}$ by

$$\tau^*(m^*(s_{(i)})) = c_{(n-i)}, \quad \textit{for} \quad i = 1, 2, ..., n-1,$$

where $\tau^*(m)$ is the optimal threshold by (9) given budget m. The optimal spending amount $m^*(s)$ is concave in intervals $(\underline{s}, s_{(1)})$, $(s_{(1)}, s_{(2)})$,...,$(s_{(n-2)}, s_{(n-1)})$, and $(s_{(n-1)}, \infty)$. At each break point $s_{(i)}$, the one-side first-order derivative satisfies $(m^*)'_+(s_{(i)}) \leq (m^*)'_-(s_{(i)})$.

Figure 3 Optimal spending amount and derivative of single-period utility



In the left panel of Figure 3, we plot an example of the optimal spending amount $m^*(s)$ with respect to s under the optimal ratio reimbursement policy. The right panel shows the corresponding marginal decrease in current period utility loss $|U_r'(m)|$ at different budget level m (recall $U_r'(m)$ is negative). The reimbursement scopes (only S or both F and S) are labeled in each panel.

The results in Theorem 3 can be interpreted as follows. With ratio policy in (i), the state space (fund level) can be divided to two regimes $(\underline{s}, s_r)$ and (s_r, ∞). In the first regime with $s < s_r$, the optimal spending amount in current period $m^*(s)$ is relatively low, and only the second-level service is reimbursed. In the second regime with $s > s_r$, the spending amount $m^*(s)$ is larger and both types of services are reimbursed. In each regime, the optimal spending amount $m^*(s)$ for current period is concave increasing in s. That is, as the fund level s increases, the marginal spending rate $(m^*)'(s)$ decreases, and the marginal savings rate $1 - (m^*)'(s)$ increases. This can be explained by the property of the per period utility loss $U_r(m)$. By Proposition 3, the marginal benefit of increasing the current period budget, $|U_r'(m)|$, decreases in the budget level m. This decreases the marginal spending rate $(m^*)'(s)$ as $m^*(s)$ increases in each regime. In particular, with $\gamma = 2$ in the power utility loss function, $|U_r'(m)|$ decreases linearly in each regime,as shown in Figure 3.

However, the concavity of the optimal policy $m^*(s)$ is not preserved over the entire state space. There is a break point at $s = s_r$, where the central planner starts to include the first-level service in the reimbursement. In addition, we can show $(m^*)'_+(s_r) \leq (m^*)'_-(s_r)$, i.e., the marginal spending rate $(m^*)'(s)$ experiences an upward jump as the fund level s exceeds s_r. This is seen in the left panel of Figure 3. It suggests that after enlarging the reimbursement scope, the central planner would increase the marginal spending rate. The behavior of $m^*(s)$ around $s = s_r$ can be interpreted by the pattern of $|U'_r(m)|$ in the right panel of Figure 3. Although the marginal benefit of increasing budget $|U'_r(m)|$ decreases in m, enlarging the reimbursement scope *slows down* its rate of decline. This motivates the central planner to increase her current period spending $m^*(s)$ more aggressively as the fund level s increases, leading to the upward jump in $(m^*)'(s)$ at $s = s_r$.

Similar interpretation applies to the cap reimbursement policy with random discrete cost levels in (ii) of Theorem 3. When the fund level s lies in the regime $(s_{(i-1)}, s_{(i)})$, the reimbursement scope is fixed: only the agents with costs $c \geq c_{(n+1-i)}$ are reimbursed, as the self-payment threshold τ^* falls between $(c_{(n-i)}, c_{(n+1-i)})$. In each of these regimes, the optimal spending amount $m^*(s)$ is concave increasing in s. When the fund level exceeds the break point $s_{(i)}$, the agents with cost level $c_{(n-i)}$ are included in the reimbursement. This slows down the decline rate in the marginal benefit of spending, leading to the upward jump in the marginal spending rate with $(m^*)'_+(s_{(i)}) \leq (m^*)'_-(s_{(i)})$.

The results in Theorem 3 provide novel managerial implications for setting the reimbursement policy in the dynamic setting. In each regime where the reimbursement scope is fixed, the central planner decreases her marginal spending rate for current period as the fund level increases. When the reimbursement scope is enlarged, the marginal spending rate experiences an upward jump, meaning that the optimal spending amount becomes more responsive to the increase in fund level. For technical reasons, we develop the results under the special case of power utility loss function with $\gamma = 2$.[5] That said, we expect the intuition to apply in other settings as well.

4.3. Horizontal Pooling in Dynamic Setting

In this section, we extend our dynamic model to multiple groups and explore the effects of horizontal pooling. We show that our main results in Section 3 still hold. We consider K groups with population weight $w^{(i)}$, services incidence rates $h_F^{(i)}$ and $h_S^{(i)}$, and random service costs $C_F^{(i)}$ and $C_S^{(i)}$ for $i = 1, 2, ..., K$. In the dynamic setting, each group manages a medical fund for its future reimbursements. At the beginning of period t, the fund levels are represented by a vector $\mathbf{s}_t := \{s_t^{(i)}\}_{i=1}^K$. In each period, group i receives a random inflow amount $q_t^{(i)}$. We assume $\{q_t^{(i)}\}_{i=1}^K$ are independent and identically distributed in different

[5] Technically, we assume $\gamma = 2$ in the utility function as the piece-wise linearity of the derivative $U'_r(m)$ is needed for our proof (see Section EC.5.5). In our numerical experiments, the piecewise concave pattern holds for other values of γ.

periods, but they can be correlated across the groups. The representative agent's utility loss function $u(l)$, interest rate r, and central planner's discount factor β are common to all groups.

With multiple groups, the central planner minimizes the total utility loss from all groups. The value function can be represented as

$$v(\mathbf{s}) := \min_{\pi} \mathsf{E}_{\mathbf{q}} \left[\sum_{t=0}^{\infty} \beta^t \tilde{u}^{(p)}(\mathbf{s}_t, \pi(\mathbf{s}_t)) \middle| \mathbf{s}_0 = \mathbf{s} \right], \tag{29}$$

where $\pi(\mathbf{s})$ is the policy function; $\tilde{u}^{(p)}(\mathbf{s}, a)$ denotes the per-period utility loss:

$$\tilde{u}^{(p)}(\mathbf{s}, a) = \sum_{i=1}^{K} w^{(i)} \tilde{u}^{(i)}(s^{(i)}, a), \tag{30}$$

with $\tilde{u}^{(i)}(s^{(i)}, a)$ defined similar to (23). To save space, we provide the explicit definitions of the state $\mathbf{s}_t$ and action a_t in the multi-group setting in Section EC.1.

We still consider three operation systems in the multigroup setting: non-pooling (NP), full pooling (FP), and monetary pooling (MP). In non-pooling, each group operates independently with its own fund and reimbursement policy. In full pooling, we pool the medical funds and inflows of all groups, and use a common reimbursement policy for them. In monetary pooling, we pool the medical funds and inflows into one, but can use different reimbursement policy for each group. The formulations of the dynamic problem under the three operation systems are given in Section EC.1. Denote the value function under the three systems by $v^{(NP)}(\mathbf{s})$, $v^{(FP)}(\mathbf{s})$ and $v^{(MP)}(\mathbf{s})$, respectively.

We compare the performance of the three operation systems in the dynamic setting. With multiple groups, the optimal policy can still be decomposed to two stages as in Lemma 2: we first determine the optimal spending amount in current period, and then set the corresponding reimbursement policy for each group. Thus, our single-period results on the performance of the three pooling systems still apply. This is summarized in the following theorem, which is the counterpart of Theorem 1 in the dynamic setting.

Theorem 4 *In the dynamic setting, full pooling and monetary pooling lead to the same optimal reimbursement policy and system performance if (i) the cap reimbursement policy is used; or (ii) the ratio reimbursement policy is used and the service costs are homogeneous across groups.*

Theorem 4 shows that our main results on the performance of the three pooling systems still hold in the dynamic setting. If the optimal cap policy is used, the full pooling system leads to the same performance as the monetary pooling system, regardless of the cost distribution. This also holds when using the optimal ratio reimbursement policy with homogeneous service costs across groups. In these scenarios, the central planner can simply pool the medical funds together and use a common reimbursement policy for all groups.

The monetary pooling only outperforms full pooling when the central planner uses the ratio reimbursement policy with heterogeneous service costs.

In the single-period setting, Remark 2 states that pooling homogeneous groups leads to no improvement. However, this conclusion does not hold in the dynamic setting, as shown by the following lemma.

Lemma 4 *Consider K homogeneous groups with same incidence rates $(h_F^{(i)}, h_S^{(i)})$, distributions of service costs $(C_F^{(i)}, C_S^{(i)})$, distribution of inflow $q_t^{(i)}$, and initial fund level $s_0^{(i)}$. In the dynamic setting, full pooling may strictly outperform non-pooling, as long as $\{q_t^{(i)}\}$ are not perfectly correlated.*

The above lemma shows that even for fully homogeneous groups, horizontal pooling can still lead to improvement in the dynamic setting. This contrasts with the single-period results in Remark 2. The reason is that in the dynamic setting, pooling multiple groups together can effectively reduce the uncertainty in the total inflow $\sum_{i=1}^{K} q_t^{(i)}$, as long as the inflows are not perfectly correlated across groups. This "inflow pooling" reveals a novel type of benefit that is absent under the single-period models. In Section EC.2.2, we show that such benefit is larger when the inflows $\{q_t^{(i)}\}$ are more volatile and/or more negatively correlated across groups. In these cases, pooling the groups together is more efficient in reducing the uncertainty of the total inflow to the pooled medical fund.

5. Numerical Study

We conduct an extensive numerical study to illustrate the managerial insights from our model. In the follows, we use the power utility loss function as in (3). We measure the system performance using the certainty equivalent loss l_{ce}. In a single period model, it is defined as the fixed net cost per capita that generates the same utility loss U of the system. In the dynamic model, we define l_{ce} as the fixed net cost per capita per period that leads to the same value function $v(\mathbf{s})$. The certainty equivalent net cost provides a straightforward and interpretable way to measure the system performance. With the power utility loss function (3), the certainty equivalent loss can be calculated as $l_{ce} = [\gamma U(m)]^{1/\gamma}$ and $l_{ce} = [\gamma(1-\beta)v(\mathbf{s})]^{1/\gamma}$ in the single-period and dynamic models, respectively.

5.1. Effects of Horizontal Pooling

We first consider horizontal pooling in the single-period setting. The optimal reimbursement policies are solved in Section 3. Without loss of generality, we assume there are two groups with the same population size. The parameters are shown in Table 3, unless they are changed in corresponding studies. In the basic setup, we set the risk aversion coefficient $\gamma = 2$; the services incidence are given by 0.5 for first-tier service (F) and 0.2 for second-tier service (S); the costs follow uniform distributions $U(2,6)$ for service F and $U(5,15)$ for service S. Thus, the second-tier service cost has larger mean and variation than the first-tier service. We consider different levels of per capita budget, as explained momentarily. Under the basic parameters, the per capita budget for fully covering the medical service costs is $\bar{m} = 4$.

Table 3 Basic Parameters in Single Period Model

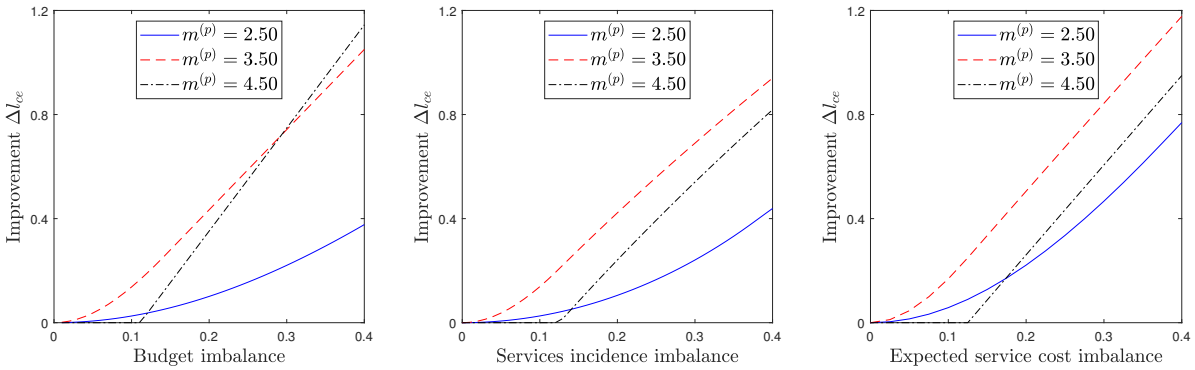
γ	$(w^{(1)}, w^{(2)})$	$(h_F^{(1)}, h_S^{(1)})$	$(h_F^{(2)}, h_S^{(2)})$	$C_S^{(1)}$ and $C_S^{(2)}$	$C_F^{(1)}$ and $C_F^{(2)}$
2	$(0.5, 0.5)$	$(0.5, 0.2)$	$(0.5, 0.2)$	$U(5, 15)$	$U(2, 6)$

We show how the benefit of horizontal pooling varies with the imbalance of the two groups and the overall budget level. We consider the imbalance in budget, service incidences, and service costs. For budget imbalance, we measure it by

$$\text{budget imbalance} = \frac{|m^{(1)} - m^{(2)}|}{m^{(1)} + m^{(2)}} = \frac{|m^{(1)} - m^{(2)}|}{2m^{(p)}}, \tag{31}$$

where $m^{(p)} := (m^{(1)} + m^{(2)})/2$ denotes the overall per capita budget level of the two groups combined. The budget imbalance lies between zero and one. A higher imbalance level indicates a larger difference between the two groups' budgets. Similarly, we measure the services incidence imbalance by $|h_j^{(1)} - h_j^{(2)}|/(h_j^{(1)} + h_j^{(2)})$ for $j \in \{F, S\}$. For service costs, we keep the distributions of $C_j^{(i)} - \mathsf{E}[C_j^{(i)}]$ unchanged, and measure the imbalance by $|\mathsf{E}[C_j^{(1)}] - \mathsf{E}[C_j^{(2)}]|/(\mathsf{E}[C_j^{(1)}] + \mathsf{E}[C_j^{(2)}])$ based on the expectations of costs. In the numerical experiments, we keep the overall budget level ($m^{(p)}$), services incidence ($h_j^{(1)} + h_j^{(2)}$), and expected costs ($\mathsf{E}[C_j^{(1)}] + \mathsf{E}[C_j^{(2)}]$) unchanged and vary the imbalance levels between the two groups. We use the same imbalance levels for the two types of services F and S.

Figure 4 Reduction in Equivalent Loss from Full Pooling (Single-Period Model)



We use the optimal ratio reimbursement policy in the numerical study. The results from the optimal cap policy are qualitatively similar. Figure 4 plots the reduction in certainty equivalent loss when we switch from the non-pooling to full pooling system, i.e., $\Delta l_{ce} := l_{ce}^{(NP)} - l_{ce}^{(FP)}$. In the first two panels, full pooling achieves the same performance as monetary pooling since the service costs are homogeneous. In each panel, we consider three overall budget levels, $m^{(p)} = 2.5$, 3.5, and 4.5. They are represented by the blue solid, red dashed, and black dashdotted lines respectively. In the middle and right panels, we keep the budget of the two groups to be the same, i.e., $m^{(1)} = m^{(2)} = m^{(p)}$.

We have consistent observations from the three panels. First, the benefit of horizontal pooling, as measured by reduction in equivalent loss, increases in the imbalance levels between the two groups. This suggests that pooling groups with higher heterogeneity in their budgets, incidence, and costs leads to larger improvement for the central planner. This reflects the nature of horizontal pooling, which is widely observed in different areas of operations management. Second, we find that in most cases, the improvement is not monotonic in the overall budget level $m^{(p)}$. In the middle and right panels, the largest improvement from pooling is always achieved under the middle overall budget level ($m^{(p)} = 3.5$). This can be interpreted as follows. When the overall budget is very low, the benefit of pooling is small as both groups are relatively poor. When the overall budget is very high, even the budget-constrained group has relatively small net cost. This limits the benefit of pooling. For example, when $m^{(p)} = 4.5$ and imbalance level is low, both groups can be fully covered even without pooling. Thus, the benefit of pooling is zero, as shown by the flat region of the dashdotted lines.

5.2. Value of Dynamic Policy

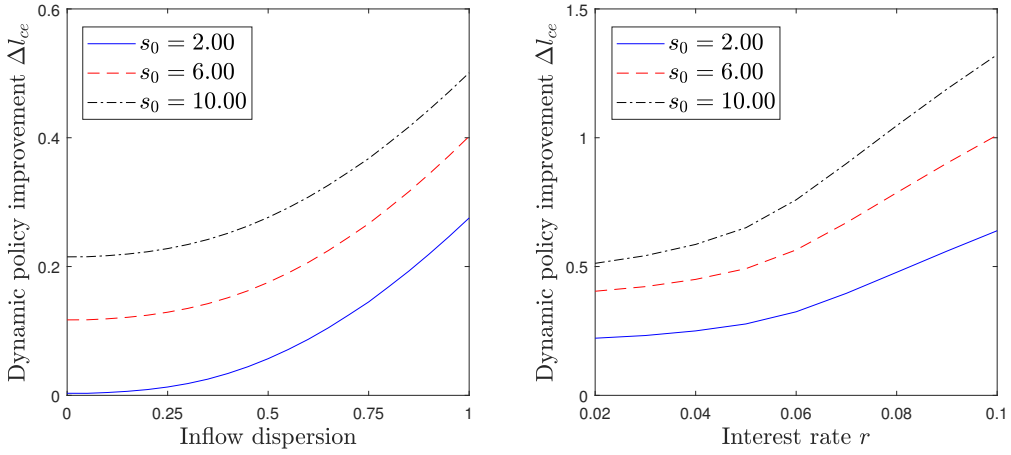
In this section, we investigate the optimal reimbursement problem in the dynamic setting. We first consider a single group with the parameters in Table 3. In the base model, the interest rate and discount factor are set as $r = 0.05$ and $\beta = 0.95$ respectively; the inflow in each period follows a uniform distribution $q \sim U(0, 3)$.

We compare the system performance from two policies. The first one is the optimal dynamic policy developed in Section 4.1, which minimizes the total discounted utility loss $v(s)$. The second one is a "myopic" policy, in which the central planner spends all the fund in current period until the services are fully reimbursed, i.e., $\pi^{(myopic)}(s) = \min\{s, \bar{m}\}$ with $\bar{m}$ defined in (7). In the myopic policy, the central planner saves for the future only after the services in current period are fully reimbursed. Comparing the system performance from the optimal dynamic policy and the myopic policy reveals the improvement from accounting for future periods in the dynamic setting. It can also be viewed as the benefit of intertemporal pooling in the multi-period model.

We measure the benefit of using the optimal dynamic policy using the reduction in the certainty equivalent loss, i.e., $\Delta l_{ce} = l_{ce}^{(myo)} - l_{ce}^{(dyn)}$. Figure 5 shows how the improvement varies with the dispersion level of inflow (left panel) and interest rate (right). We fix the expected inflow to be $\mathsf{E}[q] = 1.5$ and assume it follows a uniform distribution $U(\underline{q}, \bar{q})$. We define the inflow dispersion by $(\bar{q} - \underline{q})/(2\mathsf{E}[q])$, which takes value between zero and one. A higher dispersion level implies the inflow is more volatile. For interest rate r, we vary it from 2% to 10%, which are realistic annual levels. Three representative initial state levels s_0 are displayed.

By Figure 5, the optimal dynamic policy leads to greater improvement compared with the myopic policy when the inflow dispersion is larger and interest rate is higher. These effects are intuitive. When the per period fund inflow is more volatile, using the dynamic policy can better smooth the expenditures across

Figure 5 Reduction in Equivalent Loss from Dynamic Optimal Policy



periods, which is desirable given the convexity of the utility loss function. For example, central planner can save more when the inflow in last period is higher. On the other hand, a higher interest rate increases the benefit of savings, translating to larger improvement from the dynamic policy.

Next, we investigate the effects of horizontal pooling in the dynamic setting. We consider two groups with the basic parameters in Table 3. Their inflows are modeled as follows. We assume per capita inflow for the two groups combined follows a uniform distribution $q^{(p)} \sim U(\underline{q}^{(p)}, \bar{q}^{(p)})$. After generating a value of $q^{(p)}$, the two groups get the inflows $(1+k)q^{(p)}$ and $(1-k)q^{(p)}$, respectively. Thus, the two inflows follow uniform distributions with a correlation of one. The value of $|k|$ measures the inflow imbalance as in (31).

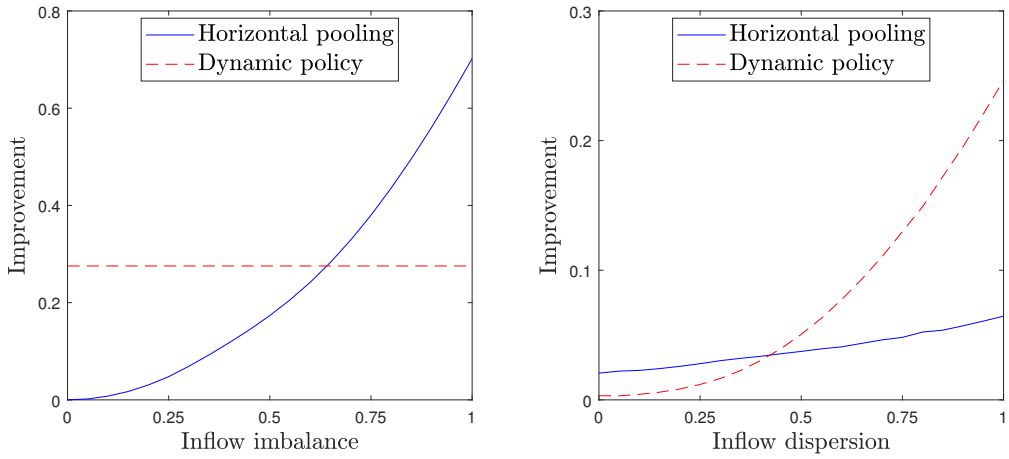
At the beginning, we assume the two groups operate independently using myopic policy. Then, the two groups are fully pooled into one and the optimal dynamic policy is used. The reduction in the certainty equivalent loss can be decomposed as following:

$$l_{ce}^{(NMP)} - l_{ce}^{(FDP)} = \underbrace{(l_{ce}^{(NMP)} - l_{ce}^{(FMP)})}_{\text{horizontal pooling}} + \underbrace{(l_{ce}^{(FMP)} - l_{ce}^{(FDP)})}_{\text{dynamic policy}}, \tag{32}$$

where $l_{ce}^{(NMP)}$, $l_{ce}^{(FMP)}$, and $l_{ce}^{(FDP)}$ denote the certainty equivalent loss under non pooling with myopic policy, full pooling with myopic policy, and full pooling with dynamic policy, respectively. The first term in (32) reflects the benefit from horizontal pooling. The second term reflects the benefit from the optimal dynamic policy, given the two groups are already pooled.

Figure 6 plots the decomposition of benefit from the horizontal pooling and dynamic policy. In the left panel, we fix the distribution of $q^{(p)} \sim U(0,3)$ and vary the imbalance level k in the inflows of the two groups (recall the two groups get inflows $(1+k)q^{(p)}$ and $(1-k)q^{(p)}$). In the right panel, we keep $q^{(1)} = 0.8q^{(p)}$, $q^{(2)} = 1.2q^{(p)}$, and the expectation $\mathsf{E}[q^{(p)}] = 1.5$. We vary the dispersion level of the inflow, which is measured by $(\bar{q}^{(p)} - \underline{q}^{(p)})/(2\mathsf{E}[q^{(p)}])$. The initial state is set at $s_0^{(1)} = s_0^{(2)} = 2$ in both panels.

Figure 6 Decomposition of Improvements from Horizontal Pooling and Dynamic Policy



As shown by Figure 6, the benefit from horizontal pooling outweighs that from optimal dynamic policy when the inflow imbalance between the two groups is high. In this case, the central planner should firstly combine the two groups horizontally via budget sharing. On the other hand, optimal dynamic policy brings greater improvement when the inflow is very volatile. This suggests the central planner should prioritize dynamic decision-making and smooth the inflows across periods. The results are qualitatively similar when we first switch to optimal dynamic policy and then pool the two groups together.

5.3. A Counterfactual Study: Pooling of China Medical Insurance Schemes

In this section, we further demonstrate the potential benefit of pooling using a counterfactual study based on real-world data. In China, most medical insurance is offered by the central government and the market share of commercial insurance is small (Blomqvist and Qian 2017). The medical insurance systems in China were highly fragmented and operated by separate authorities (Ma 2021). Different schemes are offered based on the agents' employment status, household registration system, and living locations. To provide equitable healthcare access and protect susceptible population, the authorities in China have moved towards establishing a nationwide, universal healthcare system as many developed countries. However, the progress is slow. Currently, there are two main medical insurance schemes: Urban and Rural Resident Basic Medical Insurance (URRBMI) for rural and urban residents, and Urban Employment Basic Medical Insurance (UEBMI) for urban employees.[6] The UEBMI provides a more comprehensive service coverage than URRBMI, mainly because its insurance premium is much higher. There is substantial variation in the funding source, financial protection, and benefit packages both within and across the schemes, e.g., at the county or municipal levels (Meng et al. 2015).

[6]The URRBMI itself is consolidated from the rural new cooperative medical scheme (NCMS) and urban resident-based basic medical insurance scheme (URBMI) in 2017.

We consider the counterfactual of pooling the two schemes and using a common reimbursement policy for all participants. We use the data from the National Healthcare Security Development Statistical Bulletin in 2021 and China Health Statistics Yearbook in 2022.[7] We select two representative gynecological conditions for our analysis: uterine myoma (S) and cesarean section (F). From our communication with doctors, these two conditions rarely recur within a year. In addition, their costs are relatively stable, as both are usually treated by a one-time inpatient gynecological surgery with infrequent complications. Thus, we can approximate the service incidence and costs using the average statistics reported in the yearbook. For other conditions with frequent relapsing and multiple complications, more granular data may be needed to estimate the cost distribution.

We calibrate our model parameters heuristically as follows. According to the Statistical Bulletin, the URRBMI paid \$131.14 billion for 1008.66 million participants in 2021, while the UEBMI paid \$208.03 billion for 354.31 million participants. For convenience, all amounts are converted into US dollars using the exchange rate USD/CNY $= 7.11$. The total population for the two schemes is 1362.97 million. Thus, we set the population weights as $w^{(1)} = 1008.66/1362.97 = 74.0\%$ and $w^{(2)} = 354.31/1362.97 = 26.0\%$. By the health statistics yearbook, the total number of hospital discharges is 359,144 for uterine myoma and 1,489,799 for cesarean section in 2021. We use them to estimate the services incidence for both groups as $h_F = 1.490/1362.97 = 0.109\%$ and $h_S = 0.360/1362.97 = 0.026\%$. We assume constant costs and set them at the average levels reported in the yearbook: $c_F = \$1286$ and $c_S = \$2170$. Thus, the per capita budget for full coverage is $\bar{m} = 0.026\% \times \$2170 + 0.109\% \times \$1286 \approx \$1.97$.

The yearbook does not report the specific budget allocated to the two conditions. Thus, we experiment with a range of possible budget levels as follows. First, we estimate the ratio between the per-capita budgets of the two groups using the total population and annual expenditures as $m^{(2)}/m^{(1)} = (208.03/354.31)/(131.14/1008.66) \approx 4.52$. This reflects the substantial imbalance in the two schemes. The aggregated per capita budget $m^{(p)}$ is given by $m^{(p)} = w^{(1)}m^{(1)} + w^{(2)}m^{(2)}$. Combining the two allows us to solve $m^{(1)}$ and $m^{(2)}$ from $m^{(p)}$ as:

$$m^{(1)} \approx 0.52 \times m^{(p)} \quad \text{and} \quad m^{(2)} \approx 2.36 \times m^{(p)}. \tag{33}$$

We express the aggregated per capita budget as $m^{(p)} = c_r \times \bar{m}$, where $\bar{m} \approx \$1.97$ is the level for full coverage. In China, the reimbursement ratio for medical services typically ranges from 30% to 70%. Therefore, we vary the overall coverage ratio c_r in this range to obtain possible levels of $m^{(p)}$. The budgets for the two groups $m^{(1)}$ and $m^{(2)}$ are then obtained by (33) accordingly.

[7]The bulletin is available at `https://www.nhsa.gov.cn/art/2022/6/8/art_7_8276.html`. The yearbooks are available at `http://www.nhc.gov.cn/mohwsbwstjxxzx/tjzxtjsj/tjsj_list.shtml` (in Chinese). The 2022 yearbook is the most up-to-date one, which reports the costs in 2021.

To measure the economic impact of pooling the two schemes, we translate the benefit of pooling to the equivalent relative saving in the budget level. This is measured by the constant δ_r such that $U^{(NP)}(\mathbf{m}) = U^{(FP)}((1-\delta_r)\mathbf{m})$. That is, with full pooling, the central planner can achieve the same performance for the two services as before with per capita budget decreased by α for both two groups. Then, the absolute saving in the total budget is given by $\Delta_m = \delta_r \sum_{i=1}^{2} N^{(i)} m^{(i)}$ based on the current budget level and population size. When calculating the equivalent saving, we keep the per capita budget ratio $m^{(2)}/m^{(1)}$ unchanged to reflect the imbalance between the two groups. When c_r is large, it is possible that the two conditions are already fully covered under the UEBMI scheme.

Figure 7 Equivalent Budget Saving from Full Pooling

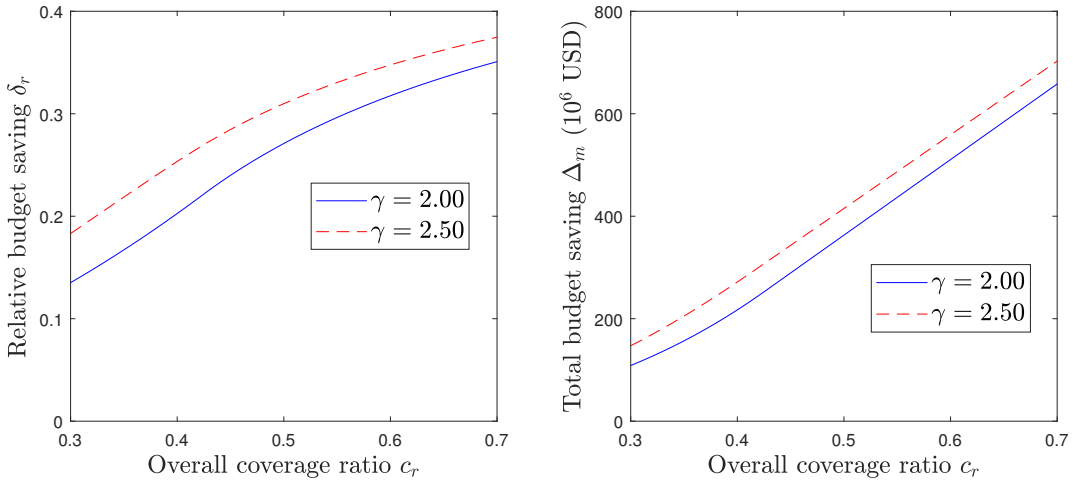


Figure 7 shows the equivalent relative saving δ_r (left) and absolute saving amount Δ_m (right) in the budget when the current coverage ratio $c_r = m^{(p)}/\bar{m}$ varies from 30% to 70%. Two risk aversion levels in the power utility loss function (3) are considered: $\gamma = 2$ by blue solid line and $\gamma = 2.5$ by red dashed line. A higher level of γ implies the central planner is more risk averse. We see that the economic effect of pooling is substantial. With $\gamma = 2$, pooling can equivalently save the budgets of two groups by 13.5% to 35.1%, translating to a total saving amount of 108.6 to 658.1 million USD. The magnitude of improvement becomes larger under a higher risk aversion level, i.e., when the central planner is more sensitive to extreme loss of agents.

The large improvement from pooling can be attributed to the substantial imbalance between the URRBMI and UEBMI schemes, each of which covers a huge population group (about 300 million and 1000 million) in China. According to the data, the per capital expenditure in the UEBMI scheme for urban employees is about 4.5 times of that in the URRBMI scheme for urban and rural residents. Such imbalance has drawn attention from researchers and authorities. Pilot actions have been taken to further integrate the two schemes,

especially in developed regions (Zhu et al. 2017). For example, Zhangjiagang City has developed a mechanism that allows residents to switch to the UEBMI scheme in the form of flexible employment (Pei 2014). These explorations are consistent with the managerial implications of our model.

6. Conclusion

In this paper, we study the design of reimbursement policy and the effects of horizontal pooling in the healthcare setting. In the model, the central planner minimizes the total utility loss by deciding the reimbursement policy for medical services, subject to a budget constraint. We find that the optimal reimbursement policy has a "max-out-of-pocket" structure, in which the central planner reimburses all the excess costs above a threshold. We also solve for the optimal ratio reimbursement policy, and show that it can be characterized based on the distributions of service costs. The model is then extended to the multi-group setting to explore the effects of pooling. Three pooling systems are considered depending on whether resource sharing and/or policy flexibility are achieved. We develop sufficient conditions such that the full pooling system can achieve the same performance as monetary pooling, suggesting that policy flexibility is not needed. Finally, we use a dynamic setting to capture the intertemporal trade-off in central planner's decision-making. The problem is formulated using a Markov decision process, and the optimal solution structure is analyzed. An extensive numerical study is performed to show the benefit of horizontal pooling, including a counterfactual study based on real-world data.

We make several contributions in the areas of healthcare operations management and health economics. We develop an analytical framework for the central planner's reimbursement policy design problem. The optimal reimbursement policies are solved explicitly, which bring managerial insights for the policy maker. Our analysis on the effects of pooling is practically relevant for the integration and consolidation of different medical insurance schemes, which have drawn significant attention in many countries. In particular, policy flexibility is more valuable when a ratio reimbursement policy is used and the heterogeneity in cost distributions is significant across groups. In the dynamic setting, we show that the central planner should increase the marginal spending rate once the reimbursement scope is enlarged, as this slows down the decline in the marginal benefit of spending.

Future research can be extended in several directions that reflect the limitations of our current work. First, it is interesting to examine the impact of using more complex reimbursement policies. For example, the central planner may use a hybrid policy that combines ratio and cap policy, or set different reimbursement ratios based on the cumulative spending amount. Second, the model may be extended to incorporate the response in patient's demand. For example, in some cases, agents may be able to choose between the two services based on their clinical outcomes and net costs. In addition, when consolidating the insurance schemes of multiple groups, the interaction and bargaining between groups may be accounted. Finally, it would be practically useful to calibrate the model with more granular data (e.g., patient-level records) and evaluate the policy impacts. These potential topics are deferred to future research.

References

Adida E, Mamani H, Nassiri S (2017) Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science* 63(5):1606–1624.

Afeche P, Pavlin JM (2016) Optimal price/lead-time menus for queues with customer choice: segmentation, pooling, and strategic delay. *Management Science* 62(8):2412–2436.

Apesteguia J, Oechssler J, Weidenholzer S (2020) Copy trading. *Management Science* 66(12):5608–5622.

Arrow KJ (1963) Uncertainty and the welfare economics of medical care. *American Economic Review* 53(5):941–973.

Ata B, Killaly BL, Olsen TL, Parker RP (2012) On hospice operations under medicare reimbursement policies. *Management Science* 59(5):1027.

Atun R, De Andrade LOM, Almeida G, Cotlear D, Dmytraczenko T, Frenz P, Garcia P, Gómez-Dantés O, Knaul FM, Muntaner C, et al. (2015) Health-system reform and universal health coverage in Latin America. *Lancet* 385(9974):1230–1247.

Baicker K, Goldman D (2011) Patient cost-sharing and healthcare spending growth. *Journal of Economic Perspectives* 25(2):47–68.

Benjaafar S (1995) Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research* 87(2):375–388.

Blomqvist Å, Qian J (2017) China's future healthcare system: What is the role for private production and financing? *International Journal of Healthcare Technology and Management* 16(1-2):29–43.

Brot-Goldberg ZC, Chandra A, Handel BR, Kolstad JT (2017) What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics. *Quarterly Journal of Economics* 132(3):1261–1318.

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

Corbett CJ, Rajaram K (2006) A generalization of the inventory pooling effect to nonnormal dependent demand. *Manufacturing & Service Operations Management* 8(4):351–358.

Cotlear D, Nagpal S, Smith O, Tandon A, Cortez R (2015) *Going universal: How 24 developing countries are implementing universal health coverage from the bottom up* (Washington, DC: The World Bank).

Culyer AJ, Newhouse JP (2000) *Handbook of Health Economics* (Amsterdam: Elsevier).

Dai T, Akan M, Tayur S (2017) Imaging room and beyond: the underlying economics behind physicians' test-ordering behavior in outpatient services. *Manufacturing & Service Operations Management* 19(1):99–113.

Deo S, Gurvich I (2011) Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science* 57(7):1300–1319.

Einav L, Finkelstein A, Cullen MR (2010) Estimating welfare in insurance markets using variation in prices. *Quarterly Journal of Economics* 125(3):877–921.

Eppen GD (1979) Effects of centralization on expected costs in a multi-location newsboy problem. *Management Science* 25(5):498–501.

Finkelstein A, Hendren N, Shepard M (2019) Subsidizing health insurance for low-income adults: Evidence from massachusetts. *American Economic Review* 109(4):1530–1567.

Finkelstein A, Taubman S, Wright B, Bernstein M, Gruber J, Newhouse JP, Allen H, Baicker K, Group OHS (2012) The oregon health insurance experiment: evidence from the first year. *Quarterly Journal of Economics* 127(3):1057–1106.

Glied S, Smith PC (2011) *The Oxford Handbook of Health Economics* (Oxford University Press).

Gross T, Notowidigdo MJ (2011) Health insurance and the consumer bankruptcy decision: evidence from expansions of medicaid. *Journal of Public Economics* 95(7-8):767–778.

Guo P, Tang CS, Wang Y, Zhao M (2019) The impact of reimbursement policy on social welfare, revisit rate, and waiting time in a public healthcare system: Fee-for-service versus bundled payment. *Manufacturing & Service Operations Management* 21(1):154–170.

Handel B, Hendel I, Whinston MD (2015) Equilibria in health exchanges: adverse selection versus reclassification risk. *Econometrica* 83(4):1261–1313.

Hartley JE (1997) *The Representative Agent in Macroeconomics* (Routledge).

Ho K, Lee RS (2019) Equilibrium provider networks: bargaining and exclusion in health care markets. *American Economic Review* 109(2):473–522.

Ikegami N, Yoo BK, Hashimoto H, Matsumoto M, Ogata H, Babazono A, Watanabe R, Shibuya K, Yang BM, Reich MR, et al. (2011) Japanese universal health coverage: Evolution, achievements, and challenges. *Lancet* 378(9796):1106–1115.

Kwon S (2008) Thirty years of national health insurance in South Korea: lessons for achieving universal health care coverage. *Health Policy and Planning* 24(1):63–71.

Lagomarsino G, Garabrant A, Adyas A, Muga R, Otoo N (2012) Moving towards universal health coverage: health insurance reforms in nine developing countries in Africa and Asia. *Lancet* 380(9845):933–943.

Liu X, Montgomery A, Srinivasan K (2018) Analyzing bank overdraft fees with big data. *Marketing Science* 37(6):855–882.

Luenberger DG (1997) *Optimization by Vector Space Methods* (John Wiley & Sons).

Ma X (2021) *Public Medical Insurance Reforms in China* (Singapore: Springer Singapore Pte. Limited), 1st edition 2022 edition.

Mandelbaum A, Momčilović P, Tseytlin Y (2012) On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science* 58(7):1273–1291.

Meng Q, Fang H, Liu X, Yuan B, Xu J (2015) Consolidating the social health insurance schemes in China: Towards an equitable and efficient health system. *Lancet* 386(10002):1484–1492.

Merton RC (1992) *Continuous-time Finance* (Cambridge, UK: B. Blackwell).

Netessine S, Rudi N (2006) Supply chain choice on the internet. *Management Science* 52(6):844–864.

Pauly MV (1968) The economics of moral hazard: Comment. *American Economic Review* 58(3):531–537.

Pei C (2014) Research on the theory and practice in the integration of urban and rural health care insurance——taking ZhangJiakou as an example. *China Health Insurance* 5:35–37.

Puterman ML (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (New York: John Wiley & Sons).

Qi A, Ahn HS, Sinha A (2015) Investing in a shared supplier in a competitive market: Stochastic capacity case. *Production and Operations Management* 24(10):1537–1551.

Sachs J, Larrain B F (1993) *Macroeconomics in the Global Economy* (New York: Harvester Wheatsheaf).

Sommers BD, Gawande AA, Baicker K (2017) Health insurance coverage and health—what the recent evidence tells us. *New England Journal of Medicine* 377(6):586–593.

Song H, Tucker AL, Graue R, Moravick S, Yang JJ (2020) Capacity pooling in hospitals: the hidden consequences of off-service placement. *Management Science* 66(9):3825–3842.

Tikkanen R, Osborn R, Mossialos E, Djordjevic A, Wharton G (2020) *International Profiles of Health Care Systems* (New York City: The Commonwealth Fund).

Wang H, Yip W, Zhang L, Hsiao WC (2009) The impact of rural mutual health care on health status: evaluation of a social experiment in rural china. *Health Economics* 18(S2):S65–S82.

WHO (2019) *Global Spending on Health: A World in Transition*. World Health Organization, Geneva.

Xu K, Chan CW (2016) Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management* 18(3):314–331.

Yang C, Hu Z, Zhou SX (2021) Multilocation newsvendor problem: centralization and inventory pooling. *Management Science* 67(1):185–200.

Yang X, Chen M, Du J, Wang Z (2018) The inequality of inpatient care net benefit under integration of urban-rural medical insurance systems in china. *International Journal for Equity in Health* 17:1–12.

Zhu K, Zhang L, Yuan S, Zhang X, Zhang Z (2017) Health financing and integration of urban and rural residents' basic medical insurance systems in China. *International Journal for Equity in Health* 16:1–8.