Optimal Differential Subsidy Policy Design for a Workload-Imbalanced Outpatient Care Network

Yewen Deng[a], Na Li[a,*], Zhibing Jiang[b], Xiaoqing Xie[c], Nan Kong[d]

[a] *Department of Industrial Engineering & Management, Shanghai Jiao Tong University, China*
[b] *Antai College of Economics and Management, Shanghai Jiao Tong University, China.*
[c] *Department of Economics and Decision Sciences, China Europe International Business School (CEIBS), China*
[d] *Weldon School of Biomedical Engineering, Purdue University, United States*

Abstract

In many countries, medical services are delivered through a multi-hospital network where a patient may have unlimited choices to different points of access to care. Due to various reasons, such a network may experience substantial workload imbalance. One way to address this challenge relies on government providing higher subsidy to incentivize patients to visit hospitals with low workload/utilization. In this research, we studied the problem of optimal government-to-patient subsidy differential (G2P-SD) policy design. We first formulated the problem with a nonlinear optimization model to minimize the total social cost (i.e., the cost of weighted wait time plus government subsidy spending) subject to the minimum workload requirement. Then we used a discrete choice model with real-world data to identify the significant influence of G2P-SD on patient hospital visit choice and numerically specified the rates of patient arrivals at a multi-hospital outpatient care network accordingly. We next developed a large-scale two-level queuing network to analyze the impact of G2P-SD on patient flows within the service network. We defined *funding efficiency* as a potential indicator to policy makers for effective budget allocation among various types of patients. Our study verified the effectiveness of modifying the G2P-SD policy, i.e., average wait time at high-workload hospitals is reduced by 26.63%, and that at low-workload hospitals is increased by 4.8%. Furthermore, our study suggested the benefit of further tailoring the policy design with consideration of influential patient attributes, which leads to a further reduction in wait time at high-workload hospitals in our Shanghai-based case study.

Keywords: Differential subsidy; Subsidy policy; Queueing network; Choice model

1. Introduction

In the United States, supported by rigorous and comprehensive outcome studies, clear guidance is basically implemented on outpatient care pathway to specify where to access care and what care

*Corresponding author
Email address: na-li03@sjtu.edu.cn (Na Li)

pathway to take for each type of patients with occurring acute conditions. For simple medical needs such as outpatient consultation for flu-like symptoms, patients usually go to see their primary care physicians at local hospitals (or local clinics, community-based hospitals/health centers). The doctors often serve as "gatekeepers" to make referrals to specialty care providers (often at general hospitals) only if the specialty care is deemed necessary from a medical point of view. To the contrary, in many other countries including China, even though local clinics are promoted by the government as the initial point of access to care, no clear guideline is implemented to standardize care access and transition. Patients needing "generalist care", such as those needing outpatient consultation for flu-like symptoms, usually prefer to visit comprehensive hospitals and experienced physicians instead [1]. On the provider side, experienced specialists, however, have to spend noticeable proportion of their worktime providing basic medical care. With shorthand notation, we refer to patients needing generalist care as GC patients.

Further, in China, the hospital system is tiered and hospitals are rated into nine tiers. Reputable general hospitals in large cities are rated on top. They receive generous financial support from the government and retain majority of the most qualified specialists. Local health centers in inland and rural areas are rated towards low. They can provide basic medical services but do not have the capacity of offering specialist care. The hospital rating exacerbates the crowding situation at general hospitals. Many Chinese patients are known to have the tendency of seeking care only at top-rated hospitals regardless of their acuity and need. As a result, longer wait times are incurred to patients in critical need of specialist care, which is evident to increase their dissatisfaction and mortality [2]. With shorthand notation, we refer to patients needing specialist care as SC patients. Meanwhile, longer worktime and over time are incurred to the providers, especially experienced and most qualified providers, which is evident to increase the incidence of their malpractice due to exhaustion [3]. On the other hand, with lower than expected service demand, local hospitals often experience low workload and low resource utilization. The phenomenon of workload/utilization imbalance is further exacerbated by the prevailing mentality among Chinese people that highly rated general hospitals can provide better care regardless of the need and by the strong mistrust on the safety and quality of care (even basic care) delivered by lowly rated local hospitals.

The challenge of workload imbalance among hospitals is known in many parts of the world, perhaps for different causes. Similar to China, many patients in South Africa go straightly to general hospitals for minor issues without prior consultation in local primary care centers [4]. In Japan, access to regional/national public hospitals is sometimes abused. Many GC patients prefer to go directly to regional/national public hospitals emergency departments rather than receiving more appropriate primary care services in community-based clinics [5]. For another example, the United Kingdom, Australia, and Hong Kong, have both free public hospitals and paid private hospitals. In these standard well-functioning two-level hospital system, patients usually visit public hospitals for government funded basic care. This has caused the overcrowding problem in public

hospitals. On the other hand, some flexibility is allowed so that patients can access basic care at private hospitals. Although patients must pay the services out of pocket, they experience little waiting time. For the list of countries with tiered hospital systems, please refer to [6]. Without loss of generality, we refer to a hospital with high workload as a HWH and a hospital with low workload as a LWH, in the remainder of the paper.

Acknowledging the workload imbalance problem experienced in many countries, there is a need to divert patient flows from HWHs to LWHs in a multi-hospital care network. One viable way to address this largely unmet need is to provide financial incentives to patients and encourage them to access care at LWHs. In this paper, we consider for the government the macro-level design problem of setting government-to-patient subsidy differentials (G2P-SDs) between HWH and LWH under fee-for-service reimbursement for outpatient care. The objective is to alleviate the issue of workload imbalance through providing patients with more subsidy at the right amount to cover their medical expenses if they consider accessing care at LWHs (in other words, incentivizing patients to visit LWHs). Meanwhile, the patient incentive design should consider the possibility of patients transferring from LWH to HWH for necessary specialist care. More specifically, we consider the setting where the government pays a higher percentage to cover the spending of medical services if received at a LWH than at a HWH. Thus through offering the subsidy differential with higher percentage on services received at LWH, the government expects under the minimum subsidy spending, more patients to choose LWH, and consequently, more balanced workload between LWHs and HWHs, less likelihood of having the overcrowding issue at HWHs and lower total social cost. For policy design, we considered a large-scale network with sizable numbers of HWHs and LWHs (e.g., entire city of Shanghai which has 163 HWHs and 1039 LWHs), and made a universal policy recommendation on G2P-SD to the government-in-charge, given patient characteristics and policy influence on patient choice between HWH and LWH. With support of a carefully crafted discrete choice experiment, we designed more precise G2P-SD policy based on not only patient age (currently effective in Shanghai) but also other individual patient characteristics. Our research is expected to provide an analytics framework for dealing with other government subsidy policy design problems as well (e.g., governmental subsidy to individuals purchasing private insurance).

The following questions were answered in our research. One, what is the influence of G2P-SD on patient choice between HWH and LWH? Two, what is the effect of patient choice on system performance? Three, what is the optimal G2P-SD to each patient class? To answer these questions, we first modeled patient choice between HWH and LWH under the influence of G2P-SD. We conducted a survey which includes a choice experiment on hospital visit for some outpatient service by assessing respondents' reactions to randomly generated G2P-SDs. We then developed a discrete choice model that incorporates various individual patient characteristics as covariates. We identified four influential attributes from the survey: age, income, medical insurance type, and preconceived outcome difference between HWH and LWH. Note that preconceived outcome differ-

ence means the difference in care outcomes between HWH and LWH preconceived by patients. To answer the second question, we built a large-scale two-level queuing network to model dichotomous patient arrivals and subsequent intra-network service processes. We demonstrated our problem-solving approach by considering a stylized Chinese hospital system at a sizeable catchment area in case studies. We developed a queueing network model and derived approximate closed-form expressions of performance measures (i.e., wait time, utilization rate, and service throughput) with respect to G2P-SD. Finally, we formulated a nonlinear program to identify the optimal G2P-SD for each patient class (defined by the covariates in the choice model) to achieve optimality in minimization of the weighted sum of HWH and LWH wait times and the government spending under minimum workload requirement limitation. Furthermore, we proposed several subsidy policy design schemes to reduce the total social cost with the limited government subsidy budget in case studies.

Similar challenges of service workload imbalance appear in many other service systems containing multiple geographically disperse service providers/stations. For example, riders in a shared vehicle system may choose to pick up and return a car (or bicycle) at a station convenient to them. This may cause the geographical imbalance problem, i.e., an excessive number of idle vehicles at some stations whereas riders can hardly rent a vehicle at some other stations (see e.g., [7], [8]). We believe such a problem can be mitigated by altering the demand arrivals through differential pricing among the stations, for which our proposed approach is also expected to be effective.

The main contributions of the paper are two folds. First, we are the first that embed discrete choice model into optimal subsidy differential pricing policy design. For choice model, we obtained real-world hospital choice experimental data from 2022 respondents in Shanghai, China, a region suffering from serious intra-network workload imbalance. The optimal subsidy differential pricing design effectively connects macro-level design of patient flow with operational-level performance measures.

Secondly, we are the first to study the China's G2P-SD policy design problem with real-world case studies. Categorically speaking, the Chinese hospital system is a two-tiered system which includes few HWHs in urban areas and economically more developed coastal regions, and many more LWHs in rural and inland areas. We considered two distinct cohorts covered by two types of medical insurance. In addition to validating the benefit of optimizing G2P-SD settings for different age groups (an age-specific policy is currently in effect in Shanghai), our case study suggested that the system performance could be further improved with consideration of income level in the policy design and through reduction of preconceived outcome difference between HWH and LWH. Moreover, we introduced the notion of *funding efficiency*, which could provide guidance on adjusting the subsidy funding among patient classes.

The remainder of this paper is organized as follows. In Section 2, we present a brief literature review on relevant research. In Section 3, we present the patient hospital choice model, the two-level

queueing network. In the section, we also present the queuing performance optimization model. Subsequently in Section 4, we illustrate our optimal subsidy differential design study through real-world cases based on Shanghai. Finally, we draw conclusions and outline future research in Section 5.

2. Literature Review

In general, our work falls in the area of healthcare subsidy policy design and analysis. In this section, we first focus on the relevant studies in this general area. Given the two distinct features in our study, we will subsequently review the literature on modeling and choice model of patient hospital visit.

We have only seen few OR/MS studies that optimize subsidy or subsidy differential to incentivize patients to visit hospitals with less workload and thus balance the workloads within a multi-hospital network. However, all of the studies applied game theory. Qu et al. [6], the work closest to ours in terms of the abstract problem setting, analyzed the interaction between a free public healthcare system with delayed care access and a paid private healthcare system which is delay-free. In their problem, patients are required to visit the public system first. To alleviate overburden of the public system, patients in the public system are offered a subsidy to use private service whenever their wait times exceed a preset threshold. The public system is modeled as an M/M/1 queue with consideration of patient wait time. The private system does not concern its wait time and is only measured by its service cost. To patients in the public system, they can observe the current wait times and decide to join or balk the system. The authors developed stylized queueing models within a game-theoretic framework, and compared various subsidy schemes in the case where patient time-sensitivity is either identical or different. Aflaki and Andritsos [9] conducted game-theoretical analysis based on M/G/1 queues for a single non-profit hospital and a single for-profit hospital. By analyzing the long-run competition between the two types of hospitals in three different system settings, the authors concluded that providing larger subsidy to the for-profit hospital causes wait time at the non-profit hospital to increase. They also concluded that in long run, providing larger subsidy to the non-profit hospital is more effective to reducing total patient costs (i.e., waiting-related and monetary) than providing larger subsidy directly to patients. Chen et al. [10] developed a mixed duopoly game to analyze the competition between a paid private service provider and a free public service provider with incorporation of patient choice between the providers. The objective of the private service provider is to maximize its expected profit, while the objective of the public is to maximize its aggregate utility. Both private and public providers are modeled as M/M/1 queues. The authors investigated the effect of offering service recipients subsidy on the objective of each provider when the service price is either regulated or not by the government.

The above papers all applied some game-theoretic approach and offered insights into the effect of subsidy policy design on improving queuing based performance measures. However, much of the analysis relies on the analytical tractability enjoyed by stylized queuing models such as M/M/1 or M/G/1. In addition, patient decision on access to care is assumed to be completely rational and it is modeled mechanistically based on the trade-off between service delay and cost. Finally, subsidy is not designed with consideration of patient demographic characteristics. These features prevent them from generating truly meaningful recommendations on health service policy and practice. In contrast, we considered a multi-hospital system with sufficiently many hospitals covering a large catchment area. In the system, hospitals differ by capacity, and the patient's balking behavior is considered. In addition, we obtained real-world choice experimental data to model patient hospital visit behavior and incorporated patient preferences in subsidy policy design.

More broadly speaking, one line of relevant research is on optimal service system design with consideration of the relationship between service price (or subsidy) and demand. Çelik and Maglaras [11] investigated this relationship in the context of revenue management. The authors assumed the demand rate is modeled as a function of the service price, and the inverse function that maps achievable demand rate into the corresponding price exists. Then the demand rate is viewed as the firm's control, with which the inverse demand function infer the price accordingly. The control problem is to maximize the expected profit by choosing admissible demand in the forms of sequencing and expediting policies. The authors derived near-optimal dynamic pricing strategies and lead-time quotation control policies.

To a lesser extent, studies related to our work also include optimal service scheduling/sequencing in a stochastic service delivery network, which has seen much more development in the context of hospital management. Several key issues are addressed in this area, including assignment of patients to appointment slots (see e.g., [12], [13]), scheduling of regular patients with consideration of randomly arrived walk-ins (see e.g., [14], [15] and [16]), design of admission control policy on whether to accept or reject an arrival or batch arrivals (see e.g., [17], [18] and [19]).

Another stream of our study is the analysis of patient hospital visit choice behavior. Discrete choice model is a behavior modeling approach that quantifies the occurrence likelihood of an individual's each possible choice from a set of mutually exclusive and collectively exhaustive alternatives. It involves parameterized utility functions in terms of observable independent variables and unknown parameters whose values are estimated from a sample of observed choices made by decision makers when confronted with a choice-making situation [20]. The approach has been extensively used in marketing, econometrics and operations management; see e.g., [20], [21], [22], [23] and [24] and the references therein. Common discrete choice models include logit, probit, and mixed logit models, among which logit model is the most widely used for its explicitly expressable and easily interpretable probability formula. Unlike other behavior modeling approaches, such as structural equation modeling [25]; system dynamics [26]; and prospect theory [27], which focus on

analyzing what factors influence behavior, a discrete choice model can specify the probability that a decision maker chooses a particular alternative. This probability is expressed as a function of observed variables that relate to the alternatives and the decision maker [28].

In health care, we have seen modeling studies on choosing which hospital to visit. Other studies focus on choices of medical service (e.g., [29],[30], [31], [32]) and physician (e.g., [33], [34]). The existing literature on hospital choice model suggests the following factors affecting patient's hospital choice behavior: provider attributes (price, equipment, capacity, etc.), patient attributes (age, gender, income, etc.), medical convenience (e.g., distance). Bronstein and Morrisey [35] developed a logistic regression model to study patient's behavior on choosing between a metropolitan hospital and a non-metropolitan hospital after deciding to bypass the nearest hospital. The authors found that travel distance and hospital equipment are important factors for rural pregnant women in Alabama. Coulter et al. [36] conducted a discrete choice experiment to investigate several factors that affect patient hospital choice, including locality, travel distance and sector of employment (i.e. public versus private). Borah [37] developed a mixed logit model to investigate patient outpatient provider choice decision and found that price and distance to the facility play the key roles. Tai et al. [38] developed a conditional logit model to study the behavior of bypassing the nearest hospital. The authors found that distance to hospital, hospital bed size, service capacity, as well as patient age, gender, marital status, are major factors among rural Medicare beneficiaries.

In addition, several studies conducted surveys on patient choice and preference in outpatient care. These studies identified various attributes of outpatient care, and mainly focused on analyzing the impact of such attributes as delay to care, flexibility of appointment times, doctor's interpersonal manner and cost of an appointment (e.g.,[39],[40],[41],[42],[43]). For example, Liu et al. [40] conducted discrete choice experiments to examine patient choice behavior in outpatient appartement scheduling. The authors examined several operational attributes, including appointment delay, doctor of choice, flexibility on appointment time and in-clinic waiting. Besides these attributes, out-of-pocket payment, which is paid by patients when seeing the doctors, is also considered. In our study, we considered an alternative payment attribute, which is presented in the form of a G2P-SD. And G2P-SD is used as a powerful incentive lever to guide patient choice in our case study. Different from the existing literature, we evaluated patient's preconceived outcome difference between HWH and LWH under the premise of informing patients in the discrete choice experiment that "doctors believe that there is no noticeable outcome difference between HWH and LWH in treating patients needing generalist care for the condition considered". Another stream of study on patient choice is by Osadchiy et al. [44]. The authors employed a general non-parametric model of patient choice to estimate patient's willingness to wait. Then they estimated the effect of wait times on patients' appointment scheduling and arrival decision. In our study, we paid attention to the impact of G2P-SD on patient choice between HWH and LWH and its impact on patient arrival and system performance. In another case, Scott et al. [45] used a discrete choice

experiment to ask parents to imagine that their children had respiratory symptoms and to choose the type of consultation. By doing this, the authors analyzed the impact of attributes, such as waiting time, who was seen, location, and whether the doctor listened, on parent's choice. Similar designs are also included in our experiments. The attribute analyzed in our study is G2P-SD, and we analyzed its effect on parent's choice of hospital for their children.

In this research, we developed a binary choice model against first-hand behavior data collected from a survey of 2022 urban and rural residents in Shanghai, China. We investigated patient choice between visiting HWH and LWH, which represent two distinct tiers in the Chinese tiered hospital system. We included the setting of G2P-SD as an independent variable in the choice model, which is a novel factor that has not appeared in the literature. By incorporating the developed choice model into optimizing subsidy differentials for the multi-hospital queuing network, our work is more comprehensive than many government subsidy policy design studies in the literature for not only considering patient insurance types, income, age, but also taking other important factors into account, such as patient's preconceived outcome difference between HWH and LWH.

3. Methodology

In this section, we first present a queueing performance based nonlinear programming model to identify optimal G2P-SD for outpatient service in a large-scale multi-hospital system with workload imbalance. This model takes into consideration patient hospital choice behavior and its impact on service system performance measures. For exposition convenience, we remove the word outpatient concerning patients who need outpatient services throughout the paper. Based on choice model, we can distinguish hospital choices for patients of different classes statistically and determine the unit-time patient volume for each class at each hospital in the network (Section 3.1). Then by applying the queueing theory, we can approximate several performance measures for the resultant queuing network model (Section 3.2). Based on these measures, we can solve the nonlinear program numerically to identify an optimal G2P-SD for each patient class.

To illustrate the service system, we consider two types of patients based on their disease conditions and specify their care pathways (see Figure 1). For patients needing specialist care, we assume they must choose an HWH to visit. For GC patients, we assume they can choose between an HWH and an LWH to visit and the choice is dependent upon the G2P-SD setting as well as many other covariates such as age group and income level. In addition, for patients of latter type choosing to visit LWH, the service received there may not suffice. As a result, some portion of these patients will be referred to HWH. Note that all patients will be considered to have the same priority of visiting HWH, since they all come with a service appointment which is scheduled based on the first-come-first-serve principle.

Based on the questionnaire used for the choice model, the cohort of respondents was devided into classes. Without loss of generality, we assume R to be the set of patient classes. For each class

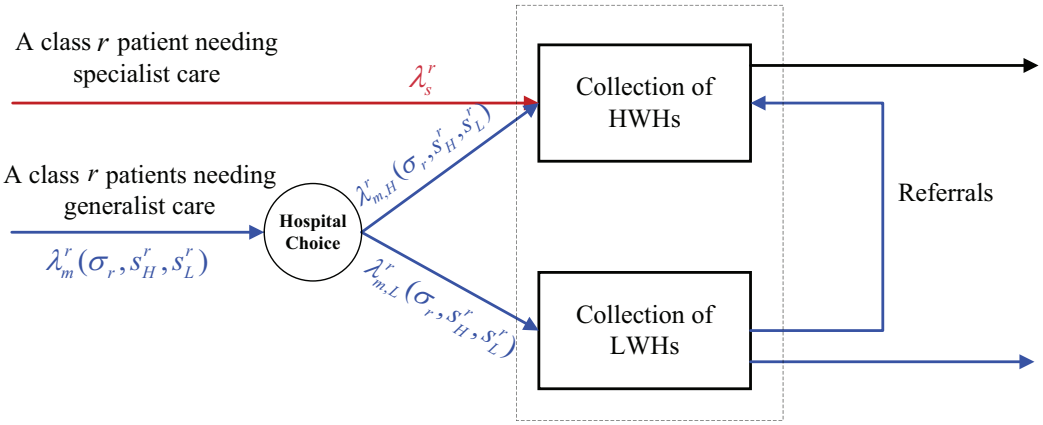


Figure 1: Illustration of the Service Flows

$r \in R$, we denote its G2P subsidy setting for HWH and LWH to be s_H^r and s_L^r, respectively. In this context, s_H^r (or s_L^r) implies that the percentage of the payment reimbursed by the government for a class r patient receiving the service at an HWH (or an LWH). We thus represent the subsidy vectors for HWH and LWH to be $\mathbf{s}_H = (s_H^1, \ldots, s_H^r, \ldots, s_H^{|R|})$and $\mathbf{s}_L = (s_L^1, \ldots, s_L^r, \ldots, s_L^{|R|})$, respectively. We denote the per-capita service payment for GC patients and SC patients by C_s and C_m, respectively. We assume that per-capita service payments at HWH and LWH are the same, e.g., both payments only involve the outpatient consultant fee that is charged identically at any hospital in China. Then the proportion of HWH payment reimbursed by the government for a class r GC patient and SC patients is $s_H^r \cdot C_m$ and $s_H^r \cdot C_s$, respectively. Similary, the proportion of per-capita payment at LWH reimbursed by the government is $s_L^r \cdot C_m$ for a class r patient.

Given the specification of patient classes, we denote $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_r, \ldots, \sigma_{|R|})$, $\sum_{r=1}^{|R|} \sigma_r = 1$, to be the class distribution vector in some catchment area (e.g., Shanghai, China). Subsequently, for each class $r \in R$, based on σ_r, s_H^r, s_L^r , one can estimate the likelihood of a patient choosing an HWH or an LWH, and then determine the exterior arrival rates. We denote $\lambda_{m,H}^r(\sigma_r, s_H^r, s_L^r)$ and $\lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r)$ to be the exterior arrival rates of class r GC patients visiting HWH and LWH, respectively; and denote λ_s^r to be the exterior arrival rate of class r SC patients visiting HWH. For notational simplicity, we further use $\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)$ and $\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)$ to represent the two $|R|$-dimensional exterior arrival rate vectors of HWH and LWH, respectively.

Given $\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)$, we denote the efficient arrival rate at an HWH i by $\tilde{\lambda}_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))$, and the mean wait time at HWH i by $W_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))$, and the utilization at HWH i by $\rho_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))$. Note that we consider the wait time for all patients as a whole because patients of different classes are given the same priority on care access. We also denote the overall throughput of class r GC patients in all HWHs as $T_{m,H}^r\left(\lambda_{m,H}^r(\sigma_r, s_H^r, s_L^r)\right)$, and that for SC patients as T_s^r which is indepedent of hospital choice. Similarly, we denote the efficient arrival rate at an LWH

j, the mean wait time at LWH j, the utilization rate at LWH j and the overall throughput of class r GC patients in all LWHs by $\tilde{\lambda}_j^L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))$, $W_j^L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))$, $\rho_j^L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))$ and $T_{m,L}^r\left(\lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r)\right)$, respectively.

Next we present the nonlinear program for the G2P-SD optimization problem.

$$\min_{\mathbf{s}_H, \mathbf{s}_L} \quad C_W \left(\alpha \sum_i^{N_H} \theta_i^H \cdot W_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) + (1-\alpha) \sum_j^{N_L} \theta_j^L \cdot W_j^L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) \right)$$

$$+ S_H + S_L; \tag{1}$$

$$\text{s.t.} \quad \theta_i^H = \frac{\tilde{\lambda}_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))}{\sum_i^{N_H} \tilde{\lambda}_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))}, \forall i \in I; \tag{2}$$

$$\theta_j^L = \frac{\tilde{\lambda}_j^L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))}{\sum_j^{N_L} \tilde{\lambda}_j^L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L))}, \forall j \in J; \tag{3}$$

$$\rho_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) \geq \rho_{min}^H, \forall i \in I; \tag{4}$$

$$\rho_j^L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) \geq \rho_{min}^L, \forall j \in J; \tag{5}$$

$$S_H = \sum_{r \in R} T_{m,H}^r\left(\lambda_{m,H}^r(\sigma_r, s_H^r, s_L^r)\right) \cdot s_H^r \cdot C_m + \sum_{r \in R} T_s^r \cdot s_H^r \cdot C_s; \tag{6}$$

$$S_L = \sum_{r \in R} T_{m,L}^r\left(\lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r)\right) \cdot s_L^r \cdot C_m; \tag{7}$$

Through adjusting the G2P subsidies, $\mathbf{s}_H$ and $\mathbf{s}_L$, our objective (1) is to minimize the cost of corresponding to the weighted sum of HWH and LWH wait times plus the subsidy spending of the government. For the weighted waiting time of the entire multi-hospital system, we assign a weight α to HWH and $1-\alpha$ to LWH. The weight α reflects the relative importance to the policy maker on achieving certain wait time oriented service level between HWH and LWH. Further, the weight assigned to the waiting time at each HWH i and each LWH j is θ_i^H and θ_j^L, respectively. And C_W is an unit-time cost of waiting for patients in the system. S_H and S_L are the unit-time total care payment at HWH and LWH funded by the government (i.e., unit-time spending on the G2P subsidies) specified in constraints (6) and (7) respectively. The objective function involves wait time, which has been a significant concern in healthcare service operations research. For example, in the study of Wan and Wang [46], the objective of implementing a subsidy scheme is to minimize the total waiting cost for all patients from the perspective of the society. Besides, the total subsidy spending in the objective function is also widely considered in healthcare service operations research. For example, in the study of Denoyel et al. [47] about the design of healthcare network under a given payment policy, the authors minimized the total cost charged to the healthcare payers. Constraints (2) indicates the proportion of the volume of patients in HWH i to the total volume of all HWHs. Constraints (3) indicate the proportion of the volume of patients in LWH j in the total volume of all LWHs. In (1)-(3), N_H and N_L represent the total numbers of HWHs and

LWHs, respectively. Constraints (4) and (5) guarantee minimum workload requirement at each HWH and each LWH, respectively.

The optimization model is a nonlinear multivariate minimization problem with continuous objective function and constraints with respect to the decision variables. Hence, we can use the Global Optimization Solver in MATLAB to obtain the solutions. The main challenge comes from how to quantify the exterior arrival rates λ_H, λ_L, and the three types of system performance measures, $W_H(\lambda_H)$, $W_L(\lambda_L)$, $\rho_H(\lambda_H)$, $\rho_L(\lambda_L)$, $T^r(\lambda_H)$, $T^r(\lambda_L)$, based on the rates. To estimate the arrival rates, we developed a binary choice model (Section 3.1). To compute the performance measures, we modeled the multi-hospital system with a large-scale two-level queuing network (Section 3.2).

3.1. Binary choice model of patient hospital visit behavior

To derive $\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)$ and $\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)$, we developed a choice model to characterize the hospital choice probabilities for different patient classes. We first designed a questionnaire that involves a choice experiment. We recruited and surveyed a cohort of online respondents residing in Shanghai. Each of the respondents was given a scenario of some imaginative common mild illness and asked about his/her hospital choice between an HWH and an LHW s/he would consider visiting. With the questionnaire, we collected individual markers (e.g., demographic characteristics) from each respondent and explained to the respondent the meaning of G2P subsidy for care payment. We then randomly assigned to the respondent a G2P subsidy on LWH within some feasibility range and informed him/her of the G2P subsidy on HWH tier currently effective. We next asked the respondent to decide between the HWH and the LWH at which s/he would make an appointment. As a result, we formed a binary choice experiment with the choice set of alternatives, denoted as $C = \{\text{HWH}, \text{LWH}\}$. With the survey, we essentially regarded each respondent a hypothetical patient so we use the term patient in the remainder of the subsection.

After completing the survey, we used a binary choice model to explore which patient-level attributes would be significant in each patient's choice behavior. We also used the model to test whether the G2P-SD setting (i.e., G2P subsidy on LWH minus that on HWH) would play an important role in the choice behavior, and how significant the effect would be. Finally, we selected a subset of independent variables given their significance in the choice model to form the patient classes.

The utilities that patient $n \in N$ chooses between HWH and LWH are given as $U_H^n = V_H^n + \varepsilon_H^n$ and $U_L^n = V_L^n + \varepsilon_L^n$. In these expressions, V_H^n and V_L^n are deterministic (representative) components, and ε_H^n and ε_L^n are random components. Each random component for the patients is assumed to be independent and identically distributed with a Gumbel distribution.

Further for the alternative HWH, we have $V_H^n = \phi_s s_H^n$, where ϕ_s is a patient-independent parameter needing to be estimated and s_H^n is the known HWH subsidy being in effect. For the

alternative choice LWH, we have $V_L^n = \phi_s s_L^n + \boldsymbol{\varphi}\mathbf{x}^n + \upsilon$, where ϕ_s is defined as above and s_L^n is the hypothetical subsidy on LWH presented to patient n. Also in V_L^n, $\mathbf{x}^n = (x_1^n, \ldots, x_k^n, \ldots, x_{|R|-1}^n)'$ is a vector of indicator variables for the patient class that patient n belongs to, via the following dummy coding scheme. For any patient n from class $r = 1$, $\mathbf{x}^n = (0, 0, 0, ...)'$; and for patient n from any other class r, only the $(r-1)^{th}$ component in $\mathbf{x}^n$ is 1 and the others are 0. Parameter vector $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_k, \ldots, \varphi_{|R|-1})$ is used to quantify the preference with respect to each patient class, which needs to be estimated. Finally, we term υ an alternative-specific constant(ASC).

For each patient $n \in I$, since ε_H^n and ε_L^n both follow Gumbel distributions, we have $\varepsilon_H^n - \varepsilon_L^n$ to be logistically distributed. Assuming patient n to be a utility maximizer, then the probability that a patient chooses the alternative LWH is given in [28] as

$$\begin{aligned} P_L\left(\mathbf{x}^n, s_H^n, s_L^n\right) &:= \Pr\left(U_L^n \geq U_H^n\right) \\ &= \Pr\left(\varepsilon_H^n - \varepsilon_L^n \geq V_L^n - V_H^n\right) = \frac{e^{V_L^n}}{e^{V_L^n} + e^{V_H^n}} = \frac{e^{\phi_s s_L^n + \boldsymbol{\varphi}\mathbf{x}^n + \upsilon}}{e^{\phi_s s_H^n} + e^{\phi_s s_L^n + \boldsymbol{\varphi}\mathbf{x}^n + \upsilon}}. \end{aligned} \tag{8}$$

Next we derive the choice model for each patient class r. Based on the earlier specification, we have $x_{l=r-1} = 1$ and the others are zero. Then given s_H^r and s_L^r, the G2P subsidies on HWH and LWH for class r patients, the probability a class r patient chooses LWH is given as

$$P_L^r\left(s_H^r, s_L^r\right) = \frac{e^{\phi_s s_L^r + \varphi_k + \upsilon}}{e^{\phi_s s_H^r} + e^{\phi_s s_L^r + \varphi_k + \upsilon}}. \tag{9}$$

Empirical evidence suggests that the subsidy rate is influential to the hospital choice and this influence differs not only by age, but income and preconceived outcome difference as well. For more detailed information, we refer the readers to Section 4.2. Therefore, SRI can be used as a viable lever to guiding patients to choose LWHs. Further, we can explore ways to incorporate these patient-specific attributes into the design of subsidy.

We assume that within every unit-time interval, there are the arrivals of K_m GC patients and K_s SC patients in the multi-hospital system. With the above definition on the LWH choice probability of patients in each class, we can derive the exterior arrival rates of LWH and HWH as

$$\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L) = \sum_{r \in R} \lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r) = K_m \sum_{r \in R} \sigma_r P_L^r\left(s_U^r, s_L^r\right) \tag{10}$$

$$\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L) = \sum_{r \in R} \lambda_s^r + \sum_{r \in R} \lambda_{m,H}^r(\sigma_r, s_H^r, s_L^r) = K_s + K_m \sum_{r \in R} \sigma_r \left(1 - P_L^r\left(s_H^r, s_L^r\right)\right). \tag{11}$$

Note that $\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L) + \lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L) = K_s + K_m$. Further, expressions (10) and (11) imply the diversion of GC patients due to individual choice. With the above exterior arrival rates, we next use them as the inputs of the large-scale multi-hospital system.

3.2. Performance evaluation for the multi-hospital system

We model the multi-hospital system as a large-scale two-level queuing network (see Figure 2). In the model, we assume N_H differently capacitated HWHs at one level and N_L differently

capacitated LWHs at the other level. Without loss of generality, we let I and J be the sets of HWHs and LWHs, respectively. For each HWH $i \in I$ and each LWH $j \in J$, we denote its capacity to be c_i^H and c_j^L, respectively. We also assume the utilization rates at HWHs are identical and those at LWHs are identical as well for tractability reason and for the reflection of the real situation in some systems. With this assumption, we have the exterior arrival rate at HWH i equipped with capacity c_i^H to be $\hat{\lambda}_i^H\left(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)\right) := \lambda_H\left(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L\right) \cdot c_i^H \Big/ \sum_{i=1}^{N_H} c_i^H$, $i \in I$ and the exterior rate at LWH j equipped with capaity c_i^L to be $\hat{\lambda}_j^L\left(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)\right) := \lambda_L\left(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L\right) \cdot c_j^L \Big/ \sum_{j=1}^{N_L} c_j^L$. For notational simplicity, we use $\hat{\lambda}_i^H$ and $\hat{\lambda}_j^L$ to present $\hat{\lambda}_i^H\left(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)\right)$ and $\hat{\lambda}_j^L\left(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)\right)$, respectively. In many tiered hospital systems, for example, the ones in urban areas in China, hospitals ranked at the same level have similar reputation and similar utilization rate. As a result, patients as a whole tend to organically distribute their hospital access across various hospitals on each tier based on the hospital's capacities.

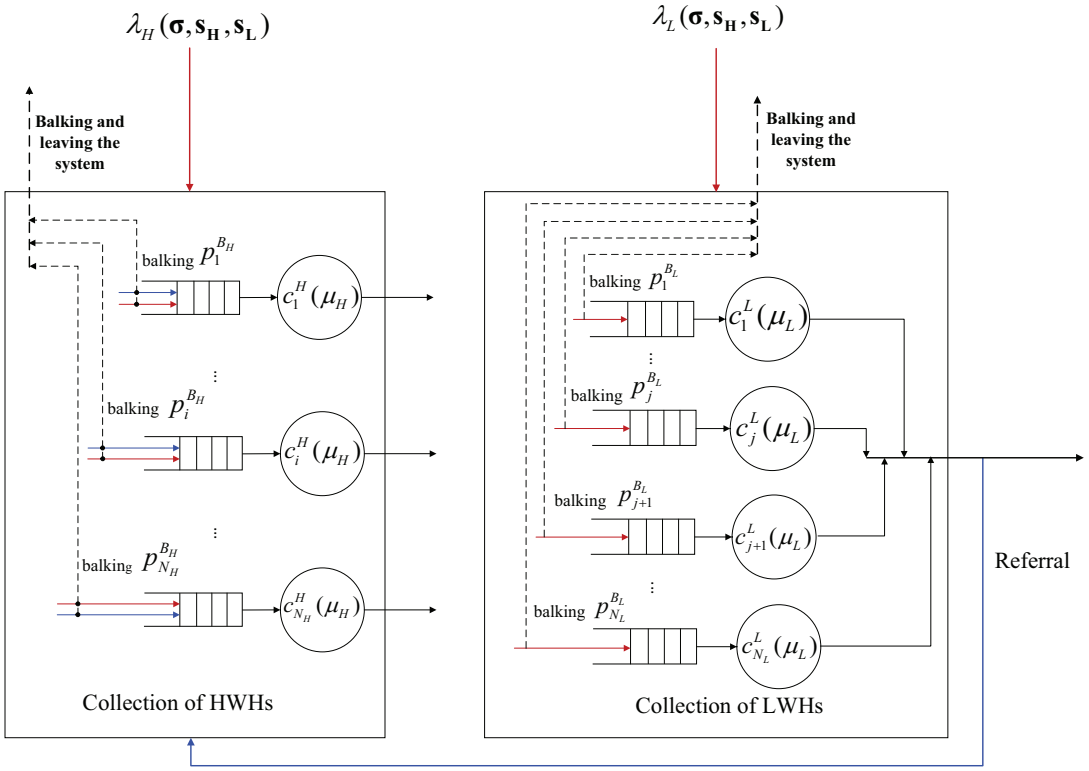


Figure 2: An illustration of the two-level queuing network with referrals

Further, we assume the arrival process at each HWH or at each LWH is a Poisson process for analysis tractability. In reality, upon arrival a patient may be discouraged by a long queue, and then s/he might refuse to join the queue (i.e., balking). We use a balking rule to consider

patient impatience and balking in the system. We first define a virtual queueing time (vqt) as introduced in [48]. The vqt is the waiting time estimated by a patient once s/he arrives at the system. An implicit assumption is that each patient can estimate the time s/he will wait from multiple sources. For example, in many countries, such as Australia, the estimated waiting time is sometimes available to patients [49], or patients may estimate the waiting time based on their previous visits [50]. Next, we introduce the balking rule. We assume an arriving patient decides to join the queue if and only if s/he evaluates that the vqt is no more than a fixed amount, which is denoted as b_H for HWH and b_L for LWH respectively. And we define the probability that vqt is more than the fixed amount b_H (or b_L) as the balking probability. The balking probability at HWH $i \in I$ and LWH $j \in J$ is represented as $p_i^{B_H}$ and $p_j^{B_L}$, respectively.

Next, we assume the service duration at each HWH and at each LWH is exponentially distributed. We denote the mean service rates of each server to be μ_H and μ_L at an HWH and an LWH, respectively. For an HWH i, an arriving patient leaves the two-level system with balking probability $p_i^{B_H}$ at his or her arrival epoch, and s/he is called a loss patient, or wait for service in an infinite capacity FCFS (first come, first served) queue at the HWH with probability $1-p_i^{B_H}$ and leave when the service is completed. Similarly, we define $p_j^{B_L}$ to be the balking probability at his or her arrival epoch at an LWH j. Finally, we consider the referrals of GC patients from LWH to HWH for not getting necessary service at LWH. We denote p_{LH} to be the LWH-to-HWH referral probability, which is attainable from the average statistics across all patients based on available LWH-HWH referral records from hospitals in the two-level system. We assume such a referral decision is made right after the exponential service process and takes place instantaneously. Similarly, referral patients from each LWH distribute their hospital access across various HWHs based on each HWH's capacity level. With the above assumptions, we formulate a two-level queueing network with each node (i.e., hospital) being an M/M/c queue with balking. Further we assume off prioritization on service access between patient classes and arrival sources.

For each HWH i, its arrival process consists of two types of arrivals: direct exterior arrivals with rate $\hat{\lambda}_i^H$, and LWH-to-HWH referrals. The total arrival rate at each HWH i is denoted by λ_i^H, $i \in I$, and the efficient arrival rate of patients (i.e., the arrival rate of patients joining the queue), denote by $\tilde{\lambda}_i^H$, is given by

$$\tilde{\lambda}_i^H = \lambda_i^H \left(1 - p_i^{B_H}\right). \tag{12}$$

For each LWH j, with the exterior arrival rate $\hat{\lambda}_j^L$, we derive the total arrival rate as $\lambda_j^L = \hat{\lambda}_j^L$ and efficient arrival rate of patients, denote by $\tilde{\lambda}_j^L$, as

$$\tilde{\lambda}_j^L = \hat{\lambda}_j^L \left(1 - p_j^{B_L}\right) = \lambda_L \left(1 - p_j^{B_L}\right) \cdot c_j^L \Big/ \sum_{j=1}^{N_L} c_j^L. \tag{13}$$

For a node in equilibrium, its arrival rate and service throughput are equal, so the overall

patient throughput of all LWHs is

$$T_{m,L} = \sum_{j=1}^{N_L} \tilde{\lambda}_j^L = \sum_{j=1}^{N_L} \left[\lambda_L \cdot \left(1 - p_j^{B_L}\right) \cdot c_j^L \Big/ \sum_{j=1}^{N_L} c_j^L \right]. \tag{14}$$

Then, with the exterior arrival rate of an HWH i, we can derive the expression for the total arrival rate and efficient arrival rate of patients at an HWH i.

$$\begin{aligned}
\lambda_i^H &= \hat{\lambda}_i^H + T_{m,L} \cdot p_{LH} \cdot c_i^H \Big/ \sum_{i=1}^{N_H} c_i^H \\
&= \frac{c_i^H \lambda_H}{\sum\limits_{i=1}^{N_H} c_i^H} + p_{LH} \frac{c_i^H \sum\limits_{j=1}^{N_L} \left[\lambda_L \left(1 - p_j^{B_L}\right) \cdot c_j^L \Big/ \sum\limits_{j=1}^{N_L} c_j^L \right]}{\sum\limits_{i=1}^{N_H} c_i^H}, \qquad (15) \\
\tilde{\lambda}_i^H &= \lambda_i^H \left(1 - p_j^{B_L}\right) \\
&= \left[\frac{c_i^H \lambda_H}{\sum\limits_{i=1}^{N_H} c_i^H} + p_{LH} \frac{c_i^H \sum\limits_{j=1}^{N_L} \left(\lambda_L \left(1 - p_j^{B_L}\right) \cdot c_j^L \Big/ \sum\limits_{j=1}^{N_L} c_j^L \right)}{\sum\limits_{i=1}^{N_H} c_i^H} \right] \left(1 - p_i^{B_H}\right). \qquad (16)
\end{aligned}$$

By Theorem 1 to Theorem 4 in [48], we can derive the probability that a patient at HWH and LWH balks, i.e., $p_j^{B_H}$ and $p_j^{B_L}$, respectively, and then the mean waiting time at the HWH and at the LWH, respectively. For more details, readers may refer to Chapter 2 of [48].

$$p_i^{B_H} = \omega_i^H \frac{c_i^H \mu_H p_{c_i^H}^H}{1 - p_{c_i^H}^H} \frac{e^{-\left(c_i^H \mu_H - \lambda_i^H\right) b_H}}{c_i^H \mu_H}, \tag{17}$$

$$p_j^{B_L} = \omega_j^L \frac{c_j^L \mu_L p_{c_j^H}^H}{1 - p_{c_j^L}^L} \frac{e^{-\left(c_j^L \mu_L - \lambda_j^L\right) b_L}}{c_j^L \mu_L}, \tag{18}$$

$$W_i^H = \begin{cases} \dfrac{\omega_i^H c_i^H \mu_H p_{c_i^H}^H \left[1 - \left(c_i^H \mu_H - \lambda_i^H\right) b_H e^{-\left(c_i^H \mu_H - \lambda_i^H\right) b_H} - e^{-\left(c_i^H \mu_H - \lambda_i^H\right) b_H}\right]}{\left(1 - p_i^{B_H}\right)\left(1 - p_{c_i^H}^H\right)\left(c_i^H \mu_H - \lambda_i^H\right)^2} & if \rho_i^H \neq 1, \\ \dfrac{\omega_i^H b_H^2}{2\left(1 - p_i^{B_H}\right)} \dfrac{c_i^H \mu_H p_{c_i^H}^H}{1 - p_{c_i^H}^H} & if \rho_i^H = 1, \end{cases}$$

$$W_j^L = \begin{cases} \dfrac{\omega_j^L c_j^L \mu_L p_{c_j^L}^L \left[1 - \left(c_j^L \mu_L - \lambda_j^L\right) b_L e^{-\left(c_j^L \mu_L - \lambda_j^L\right) b_L} - e^{-\left(c_j^L \mu_L - \lambda_j^L\right) b_L}\right]}{\left(1 - p_j^{B_L}\right)\left(1 - p_{c_j^L}^L\right)\left(c_j^L \mu_L - \lambda_j^L\right)^2} & if \rho_j^L \neq 1, \\ \dfrac{\omega_j^L b_L^2}{2\left(1 - p_j^{B_L}\right)} \dfrac{c_j^L \mu_L p_{c_j^L}^L}{1 - p_{c_j^L}^L} & if \rho_j^L = 1, \end{cases}$$

where

$$\omega_i^H = \begin{cases} \left[\frac{c_i^H \mu_H p_{c_i^H}^H}{1-p_{c_i^H}^H}\left(\frac{1}{c_i^H\mu_H-\lambda_i^H} - \frac{\lambda_i^H e^{-\left(c_i^H\mu_H-\lambda_i^H\right)b_H}}{\left(c_i^H\mu_H-\lambda_i^H\right)c_i^H\mu_H}\right)+1\right]^{-1} & if \rho_i^H \neq 1, \\ \frac{\lambda_i^H}{\lambda_i^H + \frac{c_j^L \mu_H p_{c_i^H}^H}{1-p_{c_i^H}^H}(1+\lambda_i^H b_H)} & if \rho_i^H = 1, \end{cases}$$

$$\omega_j^L = \begin{cases} \left[\frac{c_j^L \mu_L p_{c_j^L}^L}{1-p_{c_j^L}^L}\left(\frac{1}{c_j^L\mu_L-\lambda_j^L} - \frac{\lambda_j^L e^{-\left(c_j^L\mu_L-\lambda_j^L\right)b_L}}{\left(c_j^L\mu_L-\lambda_j^L\right)c_j^L\mu_L}\right)+1\right]^{-1} & if \rho_j^L \neq 1, \\ \frac{\lambda_j^L}{\lambda_j^L + \frac{c_j^L \mu_L p_{c_j^L}^L}{1-p_{c_j^L}^L}(1+\lambda_j^L b_L)} & if \rho_j^L = 1, \end{cases}$$

and

$$p_{c_i^H}^H = \frac{\left(\lambda_i^H/\mu_H\right)^{c_i^H}}{c_i^H! \sum_{n=0}^{c_i^H}\left[\left(\lambda_i^H/\mu_H\right)^n/n!\right]},$$

$$p_{c_j^L}^L = \frac{\left(\lambda_i^L/\mu_L\right)^{c_j^L}}{c_j^L! \sum_{n=0}^{c_j^L}\left[\left(\lambda_j^L/\mu_L\right)^n/n!\right]}.$$

With the balking probability at HWH and LWH, we can finally derive closed-form expressions for the relevant queueing performance measures, i.e., HWH/LWH utilization rate, mean waiting time, and class r patient throughput. That is, given N_H, N_L, $\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)$, $\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)$, $c_i^H (i \in I)$, $c_j^L (j \in J)$, μ_H, μ_L, b_H, b_L, and p_{LH}, we have

With the balking probability at HWH and LWH, we can finally derive closed-form expressions for the relevant queueing performance measures, i.e., HWH/LWH utilization rate, mean waiting time, and class r patient throughput. That is, given

- The balking probability at an HWH i: $p_i^{B_H} = \omega_i^H \frac{c_i^H \mu_H p_{c_i^H}^H}{1-p_{c_i^H}^H} \frac{e^{-\left(c_i^H\mu_H-\lambda_i^H(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)b_H}}{c_i^H\mu_H}, i \in I;$
- The balking probability at an HWH j: $p_j^{B_L} = \omega_j^L \frac{c_j^L \mu_L p_{c_j^H}^H}{1-p_{c_j^L}^L} \frac{e^{-\left(c_j^L\mu_L-\lambda_j^L(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)b_L}}{c_j^H\mu_L}, j \in J;$
- The HWH utilization rate at an HWH i:
$\rho_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) = \frac{\lambda_H(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)}{\mu_H \sum_{i=1}^{N_H} c_i^H} + p_{LH} \frac{\sum_{j=1}^{N_L}\left[\lambda_L(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\left(1-p_j^{B_L}\right)\cdot c_j^L \Big/ \sum_{j=1}^{N_L} c_j^L\right]}{\mu_H \sum_{i=1}^{N_H} c_i^H}, i \in I;$
- The LWH utilization rate at an LWH j: $\rho_L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) = \frac{\lambda_L(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)}{\mu_L \sum_{j=1}^{N_L} c_j^L}, j \in J;$

- The mean waiting time at an HWH i:
  $W_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) =$

$$\begin{cases} \frac{\omega_i^H c_i^H \mu_H p_{c_i^H}^H \left[1-\left(c_i^H\mu_H-\lambda_i^H(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)b_H e^{-\left(c_i^H\mu_H-\lambda_i^H(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)b_H}-e^{-\left(c_i^H\mu_H-\lambda_i^H(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)b_H}\right]}{\left(1-p_i^{B_H}\right)\left(1-p_{c_i^H}^H\right)\left(c_i^H\mu_H-\lambda_i^H(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)^2} & if \rho_i^H \neq 1, \\ \frac{\omega_i^H b_H^2}{2\left(1-p_i^{B_H}\right)} \frac{c_i^H \mu_H p_{c_i^H}^H}{1-p_{c_i^H}^H} & if \rho_i^H = 1, \end{cases}$$

- The mean waiting time at an LWH j:
  $W_j^L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) =$

$$\begin{cases} \frac{\omega_j^L c_j^L \mu_L p_{c_j^L}^L \left[1-\left(c_j^L\mu_L-\lambda_j^L(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)b_L e^{-\left(c_j^L\mu_L-\lambda_j^L(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)b_L}-e^{-\left(c_j^L\mu_L-\lambda_j^L(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)b_L}\right]}{\left(1-p_j^{B_L}\right)\left(1-p_{c_j^L}^L\right)\left(c_j^L\mu_L-\lambda_j^L(\boldsymbol{\sigma},\mathbf{s}_H,\mathbf{s}_L)\right)^2} & if \rho_j^L \neq 1, \\ \frac{\omega_j^L b_L^2}{2\left(1-p_j^{B_L}\right)} \frac{c_j^L \mu_L p_{c_j^L}^L}{1-p_{c_j^L}^L} & if \rho_j^L = 1, \end{cases}$$

- The respective overall throughput of class r GC patient and SC patients access HWHs

$$T_{m,H}^r\left(\lambda_{m,H}^r(\sigma_r, s_H^r, s_L^r)\right) = \sum_{i=1}^{N_H} \left[\frac{\left(\lambda_{m,H}^r\left(\sigma_r, s_H^r, s_L^r\right) + T_{m,L}^r\left(\lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r)\right) p_{LH}\right)(1-p_i^{B_H})c_i^H}{\sum_{i=1}^{N_H} c_i^H} \right]$$

$$T_s^r = \sum_{i=1}^{N_H} \left[\frac{K_s p_r (1-p_i^{B_H}) c_i^H}{\sum_{i=1}^{N_H} c_i^H} \right].$$

- The overall throughput of class r GC patient access LWHs

$$T_{m,L}^r\left(\lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r)\right) = \sum_{j=1}^{N_L} \left[\frac{\lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r)\left(1-p_i^{B_H}\right)c_j^L}{\sum_{j=1}^{N_L} c_j^L} \right].$$

We present the detailed derivations for the above performance measures in Appendix A. With these approxiamte quantities, we can substantiate the nonlinear program (1)–(7), and solve it numerically to obtain optimal patient-class-specific G2P-SDs with incorportion of the developed choice model.

4. Case Study

In this section, we use the two-tier hospital system in Shanghai, as the real-world context to present a case study. To verify our approach, we first explored the G2P-SD optimal design to achieve the minimum waiting cost and government spending. Then we compared the effectiveness of the optimal G2P-SD policy with the current policy by fixing the total government spending, in

which we carried out a series of analyses through three case studies: (1) how effective is the optimal G2P-SD design for different age groups compared to the current policy; (2) what is the benefit to the system when designing the G2P-SD policy not only based on age groups but also on income levels compared to the current policy; (3) what is the impact of the G2P-SD policy with reduced preconceived outcome difference between LWH and HWH on the system compared to the current policy. Further, we make explanations and draw conclusions about how SRIs for different patient classes should be adjusted in the G2P-SD optimal design. Through the real-world hospital choice experiment, our study is expected to offer recommendation to devise more targeted plans for both urban employee basic medical insurance and urban and rural resident basic medical insurance in Shanghai. We will start this section with a brief introduction of the Chinese hospital system and justify the need in Shanghai of differential government subsidy pricing on medical insurance policy.

4.1. Background

The 2017 Statistical Bulletin on the Development of Health Care in China (available from `www.nhfpc.gov.cn/`) reported that the average daily number of diagnosis and treatment cases handled by a primary care physician among highly rated HWHs is 7.9 as opposed to 5.7 among LWHs on the tier of LWH over entire China. From the data, we can find that the workload difference among the hospital system is obvious. Absolutely, this difference is widened in big cities like Shanghai, because care resource of higher quality is gravitated towards Shanghai than much of China. The above evidence justifies the substantial workload imbalance between the two hospital tiers.

Now we turn our attention to the situation of medical insurance in Shanghai. G2P subsidy is provided in the form of care expenses being paid partially by the government. For simplicity, we call the percentage of government-paid expenses to the total expenses the *subsidy rate*. Table 1 presents the subsidy rates for outpatient services, currently effective in Shanghai. Subsidy differentials, i.e., difference between the subsidy rates of HWH and LWH, exist for various items of basic medical care, and these differentials are age-specific in Shanghai. As described, implementation of G2P-SD is to incentivize patients to visit LWH for basic medical care, almost all of which are included in the category of outpatient care.

Table 1: Age-specific G2P subsidy rate for outpatient services covered by the two types of basic medical service insurance (UE: Urban Employee; URR: Urban and Rural Resident)

	UE insurance					**URR insurance**			
Age / **Hospital Tier**	**19-34**	**35-44**	**45-59**	**60-69**	**70+**	**0-18**	**19-59**	**60-69**	**70+**
HWH	0.5	0.5	0.6	0.7	0.75	0.5	0.5	0.5	0.5
LWH	0.65	0.65	0.75	0.8	0.85	0.7	0.7	0.7	0.7

With urban development, population aging, and personal income increase, the proportions of urban and rural resident populations in Shanghai, as well as the population distributions of age and income have changed significantly in the past decades. As a result, patient volumes from different population groups are changing. Meanwhile, we have witnessed noticeable increase of LWHs in Shanghai, i.e., the number of LWHs increased nearly 10% in the past 5 years (available from `www.stats-sh.gov.cn/`). These additional LWHs are undoubtedly changing the hospital visit behavior of some patients. In summary, it is of the government's interest to consider adjusting the G2P-SD setting accordingly. Moreover, of special interest to the context in Shanghai is the two types of basic medical service insurance: *urban employee* basic medical insurance and *urban and rural resident* basic medical insurance. For the interest of space, we refer to the two types as UE insurance and URR insurance. UE insurance covers basic medical services for employees in Shanghai, whereas URR insurance covers those for minors and Shanghai residents without job.

To estimate the aggregate arrival rates, we used the total numbers (i.e., K_s and K_m) of hourly outpatient care visits by SC patients and GC patients, respectively. The two numbers are $K_s = 1665$ and $K_m = 9435$ in entire Shanghai, according to the Shanghai Statistics Year Book (available from `www.stats-sh.gov.cn/`). Then we took the distribution between UE insured and URR insured populations and the distribution among age groups in either population, according to a survey on a sample 1% Shanghai population in 2015. Note that further distributions with respect to income and preconceived outcome difference between HWH and LWH were not available to us from a large-scale publically available dataset. We thus took the sample cohort and survey data from our choice experiment to derive the patient population distribution among more tailored classes (i.e., σ_r). Finally, with the choice probability of each class (i.e., $P_L^r\left(s_H^r, s_L^r\right)$) calculated via equation (9), we estimated the flow diversion of class r GC patients, and eventually estimated the exterior arrival rates of HWH and LWH according to equations (10) and equation (11), respectively. Note that s_H^r and s_L^r are subsidy rates of HWH and LWH for class r patients, respectively.

Through our interactions with practitioners in HWHs (e.g., Ruijin hospital, Shanghai No. 6 People Hospital) and LWHs (e.g., Xujiahui Street Community Health Service Center, Longhua Street Community Health Service Center, Hongmei Street Community Health Service Center), we were able to estimate the service rates and the referral probability. We estimated the service rates of an HWH and an LWH to be $\mu_H = 4$, $\mu_L = 4.5$, respectively, based on the average number of patients a doctor treats hourly at HWHs or LWHs. By conducting small-scale investigations in several hospitals into the threshold on the wait time tolerance, we specified the threshold at HWHs to be $b_H = 3$ hour, and the threshold at LWHs to be $b_L = 1$ hour. We estimated the probability of a GC patient being referred/transferred from LWH to HWH to be $p_{LH} = 12\%$, based on the available LWH-HWH referral records between 2015 and 2017. According to the General Hospital Classification Management Standards (available from `www.hqms.org.cn/`), HWHs and LWHs can be further divided into three categories with differences in capacity level among them. Through

an investigation of the aforementioned hospitals, we estimated that the capacity levels of HWHs in the three categories are 10, 11, 12, respectively; and the capacity levels of LWHs in the three categories are 1, 2, 4 respectively. Accordingly, we estimated the number of hospitals at each category. For HWH, the percentages of hospitals in the three categories are roughly 50%, 30%, and 20%. For LWH, the percentages are roughly 30%, 50%, and 20%. Finally, through interviews with managers of the above hospitals, we were able to obtain reasonable estimates on the costs of outpatient care for GC patients and SC patients receiving treatment in outpatient, i.e., $C_m = 300$ Chinese Yuan, $C_s = 1000$ Chinese Yuan. We write "Chinese Yuan" with "RMB", the commonly used symbol for Chinese currency, in the following.

4.2. Choice Model

To characterize patient hospital visit behaviors, we designed an online survey (see Appendix B for the web-based questionnaire). We ran the survey in May 2019 on `www.wenjuan.com`, a Chinese internet survey platform. A total of 2050 respondents participated in our study, and 2022 of which were deemed valid samples. The respondents are anonymous and the data source is reliable. The basic information collected includes gender, age, income, and insurance type. In addition, we asked about the preconceived outcome difference (POD) between HWH and LWH since it is evident that substantial workload imbalance in China can be attributed to underrating LWH and distrusting its performance among Chinese patients. Moreover, URR insurance covers minors. So we also collected information about respondents' children and asked the respondents about their hospital choices for their children. For more information on the above attributes of the correspondents, please review Table 2 (variable categorization) in the following and Table 10 and Table 11 (descriptive statistics) in Appendix D.

Next in the questionnaire, we presented a scenario. We asked each of the respondents to imagine the situation s/he experienced some common medical condition such as fever and cough with headache, muscle pain as well as other symptoms. As a result, s/he would visit some outpatient department. We then provided them with knowledge about the G2P subsidy such as its definition and current subsidy rate upon visiting the outpatient department of an HWH in Shanghai.

When presenting the scenario to each respondent, we also randomly assigned a subsidy rate for LWH within some pre-specified range to explore the effect of G2P-SD on hospital visit choice. We set the plausible range of G2P-SD from which we drew uniform samples as follows. The lower bound of the range is 0 and the upper bound is the difference between the subsidy rate of HWH and 0.95, a maximally achievable value for the subsidy rate of LWH. Since it is only meaningful to have a larger subsidy rate for LWH than that for HWH, we call the difference more specifically as *subsidy rate increment* (SRI). Note that the plausible range for SRI sampling differs by respondent's age group as the HWH subsidy rate varies by age with the current UE insurance policy (see Table 1). Table 2 presents the SRI sampling ranges.

Table 2: Attribute levels for participants of UE and URR insurance types

Insurance Type	Attribute	Level
UE	Age (yrs)	19–34, 35–44, 45–59, 60+
	Income (RMB)	[0,3K), [3K,5K), [5K,10K), [10K,30K), [30K,+∞)
	POD	No difference, relatively small, relatively large, very large
	SRI	19–34 years old: continuous in [0,0.45] 35–44 years old: continuous in [0,0.45] 45–59 years old: continuous in [0,0.35] over 60 years old: continuous in [0,0.25]
URR	Age (yrs)	0–18, 19–59, 60–69, 70+
	Income (RMB)	[0,3K), [3K,5K), [5K,10K), [10K,30K), [30K,+∞)
	POD	No difference, relatively small, relatively large, very large
	SRI	0–18 years old: continuous in [0,0.45] 19–59 years old: continuous in [0,0.45] 60–69 years old: continuous in [0,0.45] over 70 years old: continuous in [0,0.45]

With the behavior data, we parameterized a binary choice model for each insurance type. Table 3 presents our choice model results, i.e., respective estimates of parameters ϕ_s, $\boldsymbol{\varphi}$ and v, from equation (8). We found that SRIs of both insurance types are significant and positive, which suggests increasing the increment on subsidization of LWH as opposed to HWH would result in more willingness to visiting LWH. This observation strengthened our belief that creating sufficiently large SRIs can be a viable approach to providing financial incentive and to guide patients to choose LWHs (i.e., the right hospitals) and thus alleviate system workload imbalance. Our results also suggest that age is a factor as significant as SRI, especially for older adults. Further, the results imply when other factors being equal, older adults tend to be more willing to visit LWH. These results promoted us to explore the interplay between age and SRI and refine the current age-specific G2P-SD policy to further improve the system performance. Similarly, we found that income is a significant factor , except low-income groups. Our results further imply when other factors being equal, patients with lower income tend to be more willing to visit LWHs. On the other hand, patients under the coverage of UE insurance have stable income and thus are less sensitive to out-of-pocket care spending. Hence we elected to examine the effect of different income levels on the system performance and further tailor the G2P-SD policy design based on the combination of age group and income levels. Finally, our results suggest preconceived outcome difference between LWH and HWH plays a significant role. That is, the worse outcome of LWH preconceived by some patient compared to HWH, the less likely the patient would choose to visit LWH. Thus we elected to explore the impact of hypothetically configured distributions of preconceived outcome difference on the system performance, which offers insights into designing educational campaigns

among consumers on similar outcomes from basic medical care.

Further, we analyze the interaction effects between patient-specific variables and SRI, we provided our binary choice model estimation results in Table 9 in Appendix C. The results show that for the UE insurance type, age, income, and POD significantly interact with SRI. For the URR insurance type, the interaction effects of SRI with age and POD are significant. For more detailed information, we refer the readers to Appendix C.

Table 3: Choice model coefficient estimation results

Insurance Type	**Variable Name**	**Parameter Estimate**	**Standard Error**	$\boldsymbol{p}$ **−Value**
UE	SRI	7.534	0.701	0.000
	Age_{$35-44$}	0.391	0.196	0.046
	Age_{$45-59$}	0.636	0.200	0.001
	Age_{60+}	1.065	0.290	0.000
	Income_{$3K-5K$}	-0.229	0.460	0.619
	Income_{$5K-10K$}	-1.040	0.451	0.021
	Income_{$10K-30K$}	-1.719	0.476	0.000
	Income_{$30K+$}	-3.277	0.232	0.008
	POD_{relatively small}	-0.568	0.210	0.007
	POD_{relatively big}	-2.134	0.237	0.000
	POD_{very big}	-2.266	0.380	0.000
	ASC	-0.117	0.501	0.815
URR	SRI	2.292	0.372	0.000
	Age_{$19-59$}	0.952	0.103	0.000
	Age_{$60-69$}	1.229	0.274	0.000
	Age_{70+}	2.000	0.277	0.000
	Income_{$3K-5K$}	-0.169	0.144	0.238
	Income_{$5K-10K$}	-0.246	0.140	0.079
	Income_{$10K-30K$}	-0.337	0.181	0.037
	Income_{$30K+$}	-1.053	0.413	0.002
	POD_{relatively small}	-0.418	0.135	0.000
	POD_{relatively big}	-1.581	0.150	0.000
	POD_{very big}	-2.359	0.249	0.000
	ASC	-0.081	0.188	0.665

In summary, with the choice model, we identified two key factors in addition to age on patient's hospital visit behavior, which promoted us to study various G2P-SD policy redesign issues (see Study 1 and Study 2.1 – 2.3 in Section 4.3).

4.3. Results

In this section, we report four Shanghai-based case studies. We stated the four research questions at the beginning of Section 4. Our main results are subsidy rate increments (SRIs) from HWH to LWH. These studies involve solving the G2P-SD optimization model, i.e., Eq. (1) – (7), with the parameterization described in Section 4.2. Further, we set α, the weighting coefficient

between the waiting times at HWH and LWH (i.e., W_H vs. W_L), to be 0.95, for paying more attention on wait times at HWH at this proof-of-the-concept stage. And we set C_W, the unit-time cost of waiting, to be 3.0×10^6.

In the following studies, for exposition simplicity, we refer to the subsidy rate increment from HWH to LWH *in effect* as SRI_IE; the subsidy rate increment obtained from solving the nonlinear program as SRI_REC or SRI_REC_X, i.e., *recommended* subsidy rate increment. For the following studies, we considered a hypothetical cohort combining patients covered by the two types of insurance, and a separate choice model was developed for each insurance type first.

Study 1: What is the optimal G2P-SD design to achieve the best system performance?

In this study, patients covered by each insurance type are only divided into four age-specific classes; referring back to Table 2 for detailed information. We solved the G2P-SD optimization model to obtain the optimal SRIs (i.e., SRI_REC_0) for different age groups. We also compute the corresponding actual spending and wait times of HWHs and LWHs.

Figure 3 represents the comparison between the two SRI settings with respect to the insurance types and different age groups. The figure shows an increase on SRI for the UE-type insurance (e.g., for age group 19 – 34, SRI is around 15% under the SRI_IE setting vs. around 30% under the SRI_REC_0 setting), and decrease on SRI for the URR-type insurance across all four age groups (i.e., 20% under the SRI_IE setting vs. around 8% under the SRI_REC_0 setting). To evaluate the performance of G2P-SD policy under the SRI_REC_0 setting, we compared the total social cost (i.e., the cost of weighted wait times plus the government subsidy spending) under the two SRI settings (SRI_IE vs. SRI_REC_0). The result shows a 55.99% reduction in the total social cost under the SRI_REC_0 setting, compared to that under the SRI_IE setting (i.e., 7.30×10^6 under the SRI_IE setting vs. 3.21×10^6 under the SRI_REC_0).

Further, figure 4 represents the comparison on the actual spending at all hospitals. The figure shows a 9% increase in the actual spending under SRI_REC_0 compared to that under SRI_IE. Next, Figure 5 shows the comparison on the waiting times at HWHs and LWHs under the two SRI settings (SRI_IE vs. SRI_REC_0). The figure shows the wait time at HWHs has been significantly reduced under SRI_REC_0. Although the wait time at LWHs has increased to some extent, the wait time after the increase is still tolerable. The results indicate that an appropriate increase in the subsidy expenditure, especially the increase in SRI for UE patients, can effectively reduce the wait time at HWHs and balance the workloads of HWHs and LWHs.

Next, to summarize the quantifiable implications from the specifications of SRI_REC_0, we use Table 4 to introduce the notion of *funding efficiency* to verify the consideration on government funding allocation between the two types of insurance. The quantity is calculated as follows. Let us denote ΔS to be the relative change in the actual spending on all patients before and after the SRI change for a given patient class. Similarly, let us denote ΔO to be the relative change in

weighted average wait time (i.e., objective function value) when the SRI is changed for the given patient class. Then *funding efficiency* for that particular class is $\Delta O/\Delta S$. *Funding efficiency* essentially reflects the cost-benefit potential of making subsidy budget reallocation. As reported in Table 4, we compared the *funding efficiency* values for overall UE-type and URR-type insured patients. We found that the efficiency of funding for UE insurance patients is higher than that for URR insurance patients across the board. This is explained that SRIs for UE-type insurance have increased. Further, based on the comparison of this quantity among different age groups for the same type of insurance, It is not difficult to find that the greater the cost-benefit potential, the greater the increase in their SRI, which is consistent with the conclusion draw from Figure 3. We thereby recommend *funding efficiency* as an indicator for optimal SRI policy design.

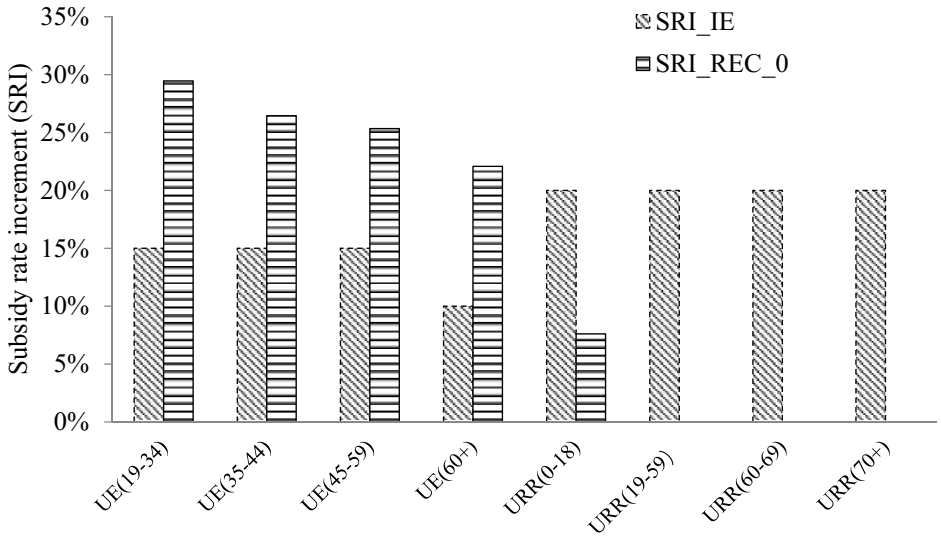


Figure 3: SRI comparison with respect to age group for UE and URR insurance types

Finally, we further explored how the effectiveness of the optimal G2P-SD policy will be affected if the government subsidy budget is limited. We first introduced a parameter S_B as the budget limit on the government subsidy spending. Thus, a budget constraint (i.e., $S_H + S_L \leq S_B$) is introduced in the G2P-SD optimization model. By solving the optimal model accordingly, we explored the impact of increasing government subsidy budget (i.e., S_B) on the performance of the optimal G2P-SD policy shown in Table 5. Table 5 reports the relative change in total social cost, the relative change in average wait times at HWHs and at LWHs, and the actual spending under the increasing subsidy budget limit. As the subsidy budget increases, the general trend of the total social cost, the HWH wait time, and the actual spending is decreasing, and the LWH wait time increases. With further subsidy budget increase, we noticed that the actual spending to achieve the optimal system performance stops increasing after the subsidy budget increases to

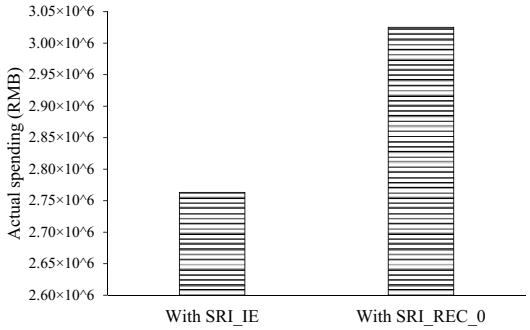


Figure 4: Actual spending under the SRI_IE and SRI_REC_0 setting

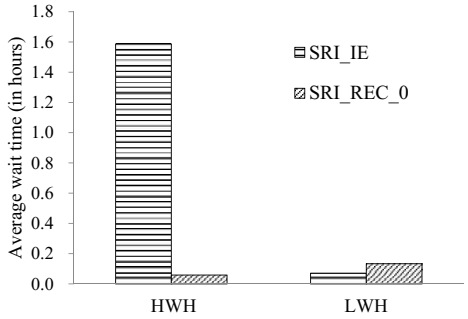


Figure 5: Comparison of the average wait times

Table 4: Changes in the cost of weighted average wait time and actual spending with 5% SRI change

Insurance type	Age group	**The cost of weighted wait time**			**Actual spending**			
		Relative change*	Before the change**	After the change***	Relative change*	Before the chage**	After the change***	Efficiency
UE	19-34	-60.1%	4534052	1810223	2.1%	2763185	2820956	28.73
	35-44	-27.7%	4534052	3278884	1.0%	2763185	2791291	27.22
	45-59	-36.0%	4534052	2900228	1.3%	2763185	2800297	26.83
	60+	-32.2%	4534052	3072946	1.1%	2763185	2794855	28.12
URR	0-18	-0.8%	4534052	4496780	0.0%	2763185	2764530	16.88
	19-59	-5.5%	4534052	4286577	0.4%	2763185	2774361	13.50
	60-69	-1.0%	4534052	4490851	0.1%	2763185	2765357	12.12
	70+	-0.6%	4534052	4506230	0.1%	2763185	2765122	8.75

* The change (either in the cost of weighted average wait time or actuall spending) when comparing the system before the SRI changed to that after the SRI changed divided by the corresponding value under SRI_IE

** The value (either the cost of weighted average wait time or actual spending) before the SRI of a patient class is changed

*** The value (either the cost of weighted average wait time or actual spending) after the SRI of a patient class is changed

certain level, thus the total social cost reaches the optimal.

Study 2: How to improve the current policy based on the cases in Shanghai?

To compare the effectiveness of the optimal G2P-SD policy without increasing the total government spending with the current policy, we took the actual government spending under the SRI_IE setting as the budget limit and used it in Study 2.1–2.3. We first estimated the actual government spending under the SRI_IE setting via Eq. (6) – (7), and assigned it to S_B. By solving the optimal model considering the budget constraint, the government subsidy budget is reallocated to obtain the optimal SRIs (i.e., SRI_REC_X) for different patient classes and the corresponding wait times

Table 5: The effectiveness of optimal G2P-SD policy under increasing subsidy budget

Subsidy budget (RMB)	Relative change in total social cost*	Relative change in average wait times at HWHs*	Relative change in average wait times at LWHs*	Actual spending (RMB)
2763200**	-16.50%	-26.64%	4.69%	2763200
2813200	-43.27%	-70.95%	17.83%	2813200
2863200	-51.97%	-86.13%	33.42%	2863200
2913200	-54.69%	-91.66%	49.56%	2913200
2963200	-55.69%	-94.41%	66.13%	2963200
3013200	-55.98%	-96.03%	83.17%	3013200
3025154***	-55.99%	-96.32%	87.32%	3025154
3063200	55.99%	-96.32%	87.32%	3025154
3113200	55.99%	-96.32%	87.32%	3025154

* The change (either in the total social cost or weighted average wait times) when comparing the system under the optimal G2P-SD policy to that under SRI_IE divided by the corresponding value under SRI_IE

** The initial value of the budget limit is set to the actual government spending under SRI_IE

*** The budget limit is set to be the government spending under SRI_REC_0

at HWHs and LWHs. Then under the same government subsidy budget, we can intuitively compare the wait time under the SRI_IE setting with that under the SRI_REC_X setting in different scenarios.

Study 2.1: How much improvement can the G2P-SD optimal design bring to the system performance compared to the current policy under the same government subsidy budget?

For this study, patients with each insurance type are also divided into four age-specific classes like Study 1. By considering the government subsidy budget limit obtained under the SRI_IE setting, we solved the G2P-SD optimization model to obtain the optimal SRIs for the four age-specific classes (i.e., SRI_REC). Meanwhile, we analyzed the queuing network to obtain the mean wait time of each hospital under the SRI_IE setting and under the SRI_REC setting respectively, and estimated the weighted average of wait times of HWHs and LWHs via Eq. (1).

Figure 6 presents the comparison between the two SRI settings with respect to the two insurance types and different age groups. The figure shows a decrease on SRI for the URR-type insurance across all four age groups (i.e., 20% under the SRI_IE setting vs. 0% under the SRI_REC setting). For the UE-type insurance, three of the four age groups see an increase on SRI (i.e., age groups 19 – 34, 35 – 44, and 60+). Moreover, the increase is more noticeable for youngest and oldest age groups than age group 35 – 44. On the other hand, there is a decrease in age group 45 – 59. By comparing the total social cost under the two SRI settings (SRI_IE vs. SRI_REC). The

result shows a 16.50% reduction in the total social cost under the SRI_REC_0 setting, compared to that under the SRI_IE setting (i.e., 7.30×10^6 under the SRI_IE setting vs. 6.09×10^6 under the SRI_REC).

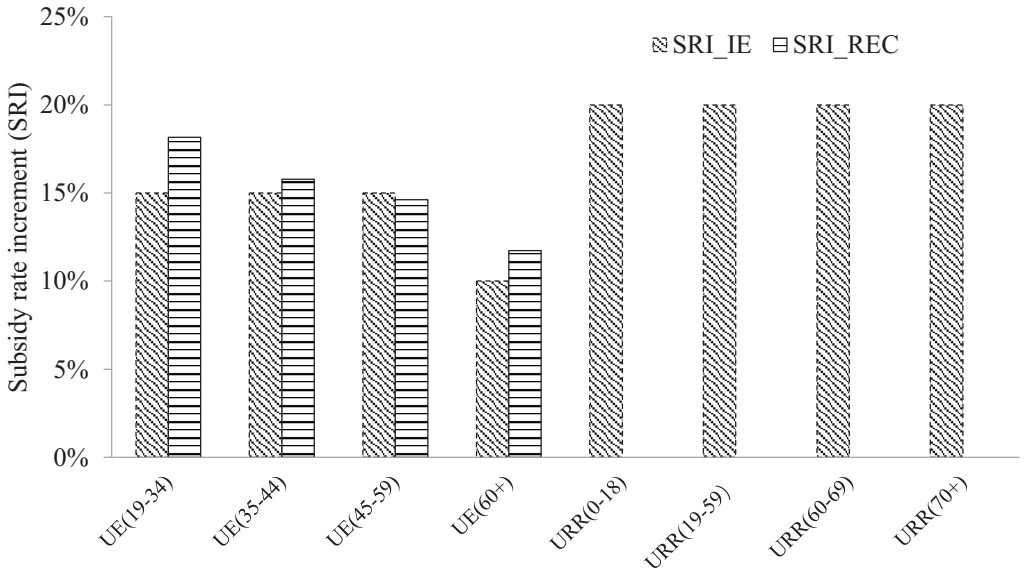


Figure 6: SRI comparison with respect to age group for UE and URR insurance types

Moreover, Figure 7 presents the comparison on the proportions of patient volume and actual spending at the hospitals on the LWH tier as opposed to the total patient volume and subsidy spending, which are two constants. In addition, to some degree, this figure implies the consequence of the SRI changes shown in Figure 6. With widened subsidy differentials for the UE-type insurance (at least the younger and older age groups), clearly more patients would choose to visit an LWH given the positive coefficient on SRI (the point estimate is 7.534). Consequently, more patients would lead to increased spending at LWHs. On the other hand, with narrowed subsidy differentials for the URR-type insurance, clearly fewer patients would choose to visit an LWH given the positive coefficient on SRI (the point estimate is 2.292). Consequently, fewer patients would lead to decreased spending at LWHs.

Next, Figure 8 presents the comparison on the weighted waiting times at HWHs and LWHs under the two SRI settings (SRI_IE vs. SRI_REC). The figure shows a 26.63% reduction in mean wait time at HWHs and a slight increase at LWHs under SRI_REC. The results suggest we have provided encouraging evidence to the hospitals on the effectiveness of reallocating the subsidy budget for reducing HWH congestion and balancing the workloads on the two tiers. It is important to have the buy-in from the multi-hospital system, which in China is a so-called hospital alliance where an HWH tends to be the leader, thus having much more bargaining power than partnering LWHs.

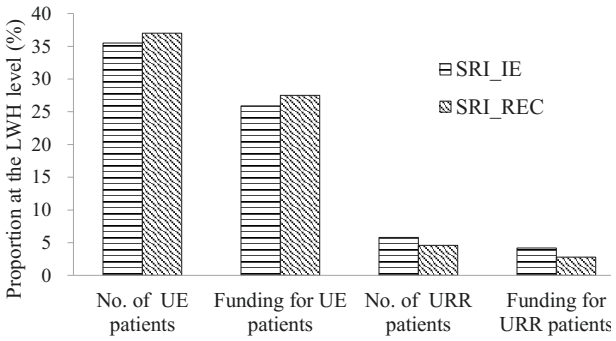


Figure 7: Comparison of the proportions at LWHs

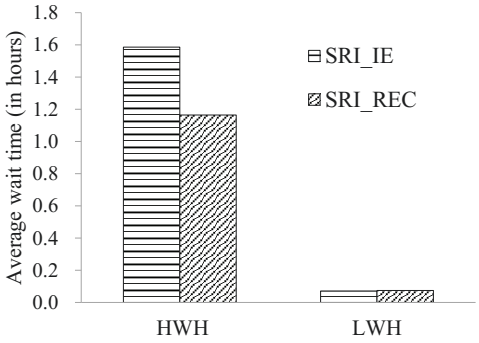


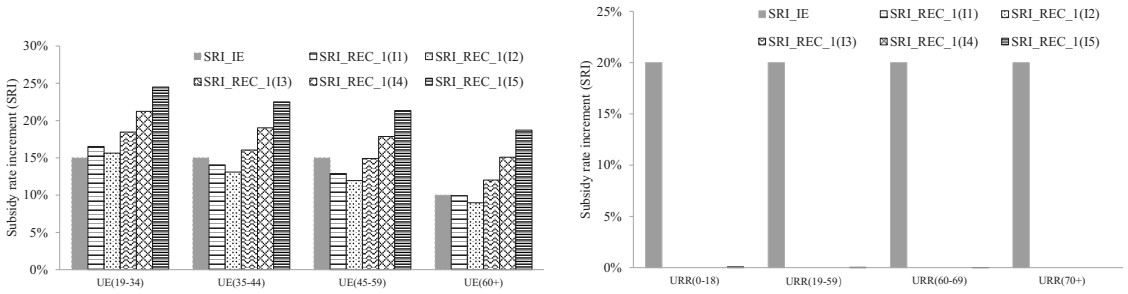
Figure 8: Comparison of the average wait times

Then, we use Table 4 to summarize the quantifiable implications of reallocating the fixed government funding according to the specifications of SRI_REC. According to Study 1, the efficiency of funding for UE insurance patients is higher than that for URR insurance patients. It is also known that there is a fixed amount of government spending to allocate. This quantity helps explain why a larger (smaller) proportion of funding goes to UE (URR) patients, which is backed by a widened (narrowed) SRI being presented to UE (URR) patients. Further, based on the comparison of this quantity among different age groups for the same type of insurance, we drew the same conclusion as from Figure 6.

Study 2.2: What is the benefit of considering additional patient characteristics in the G2P-SD design?

For this study, we compared the effectiveness of the G2P-SD policy that considers additional patient characteristics with the effectiveness of the current policy by fixing the government spending obtained under the SRI_IE setting. We further divided the patient classes based on income level, which appeared to be an influential secondary attribute from the choice model (see Table 3). As a result, we had 20 distinct age- and income-specific classes (i.e., ten for each insurance type). After a similar setup as in study 2.1, we obtained the optimal SRIs (i.e., SRI_REC_1) for different age and income-specific classes and the corresponding mean wait times of HWHs and LWHs.

Figure 9 presents the comparison between the two SRI settings with respect to the two insurance types and different patient classes. Each index in parentheses in the figure legend indicates a level of income in RMB, i.e., I1: $<$ 3K; I2: 3K-5K; I3: 5K-10K; I4: 10K-30K; I5: $>$30K. The figure suggests further tailoring the SRI policy across different income levels. For the UE insurance type, the results suggest that at the same income level, the trend of changes in SRI across age groups is consistent with the results of study 1, and bigger increases in SRI are set for the highest, the second highest, and the third highest income populations. In return, the middle-income groups should be given smaller SRI. For the URR insurance type, the results suggest significantly decreasing SRIs across all age groups and income levels. We compared the total social cost under the three SRI



(a) UE insurance (b) URR insurance

Figure 9: SRI comparison with respect to age-income group for UE and URR insurance types

settings (SRI_IE vs. SRI_REC vs. SRI_REC_1). The results show that the total social cost under the SRI_REC_1 setting is reduced by 17.59% and 1.3%, respectively, compared with the SRI_IE setting and SRI_REC setting (i.e., 7.30×10^6 under the SRI_IE setting vs. 6.09×10^6 under the SRI_REC vs. 6.01×10^6 under the SRI_REC_1).

Next, Figure 10 presents the comparison on the mean wait times at HWHs and LWHs under the three SRI settings (SRI_IE vs. SRI_REC vs. SRI_REC_1). The figure shows a further reduction (28.39%) in wait time at HWHs and a tiny further increase at LWHs under SRI_REC_1. This study further promotes the usefulness of patient hospital visit choice behavior studies. The results give encouraging evidence to the hospitals on the effectiveness of further tailoring the SRI policy. It should be noted that, although the policy performs well for improving the system performance, ethical issues involved should be carefully considered before the policy implementation.

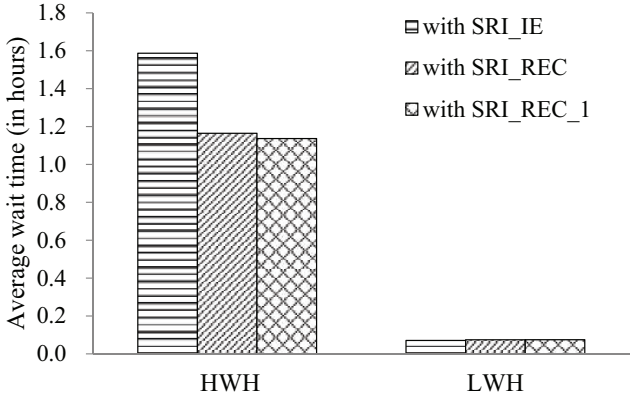


Figure 10: Comparison of the average wait times

Finally, we use Table 6 to illustrate the usefulness of *funding efficiency* as an indicator to optimal SRI policy design. After reviewing Table 6 and Figure 9, we concluded that there is clear correspondence between the value of funding efficiency and the SRI. For example, for the

UE insurance type, the largest value appears in the patient class of age group 19 – 34 and income level > 30K, we see the most widened SRI occurs in this class. On the other hand, for the URR insurance type, the value is smaller than that for the UE insurance type, we see the most narrowed SRI are set for all patient classes.

Table 6: Funding efficiency

Insurance type	Age group (yrs)	Income level (RMB)				
		<3K	3K–5K	5K–10K	10K–30K	>30K
UE	19–34	26.1	26.0	28.1	28.8	29.7
	35-44	24.6	24.2	26.2	27.5	28.9
	45-59	24.1	23.6	25.6	26.9	28.3
	60+	25.1	24.7	26.8	28.2	29.7
URR	0–18	16.6	17.1	17.4	17.7	19.5
	19-59	12.8	13.5	13.8	14.2	16.9
	60-69	11.5	12.3	12.6	13.0	15.9
	70+	8.1	8.9	9.2	9.6	12.8

Study 2.3: What is the impact of G2P-SD optimal design with reduced difference on preconceived LWH outcome?

This study was inspired by the well-documented assertion that many people in China do not trust local and community-based hospitals. As mentioned earlier, this mistrust has excerbated the situation of workload imbalance in the tiered Chinese hospital system. In this study, we analyzed the impact of G2P-SD optimal design with reduced preconceived outcome difference between LWH and HWH on the system compared to the current policy. The comparison was made with the same government spending obtained under the SRI_IE setting.

Data from our survey support this assertion. For example, our data show that with either insurance type, more than 39% of the online respondents preconceived large difference in care outcmes between the two tiers. Nevertheless, many healthcare professionals in China believe the actual care outcomes at LWH are not as bad as what are preconceived by the consumers. The interactions with our partners in Shanghai indicated this as well. Hence, it would be beneficial to reduce the preconceived outcome difference (POD), or in other words, correct the biased perception among Chinese consumers. The quesiton that remains is how much impact it would be with any hypothetic reduction of POD. Answering this question can help local governments investigate the trade-off between financially incentivizing consumers and using promotional campaigns to correct the quality perception on LWH. If found beneficial with the latter approach, our approach can further help select cost-effective public campaign strategies to implement.

In this study, we varied the POD distributions for the two types of insurance (see Figure 11 for a comparison of the two distributions). We essentially moved a portion of the consumers having relatively large POD to having relatively small POD, which is aligned with the belief of

the healthcare professionals we interacted with. In the same study setup as in Study 2.1 except for the POD distribution, we considered four age-specific patient classes and obtained the optimal SRI (i.e., SRI_REC_2) for each class and the corresponding wait times of HWH and LWH.

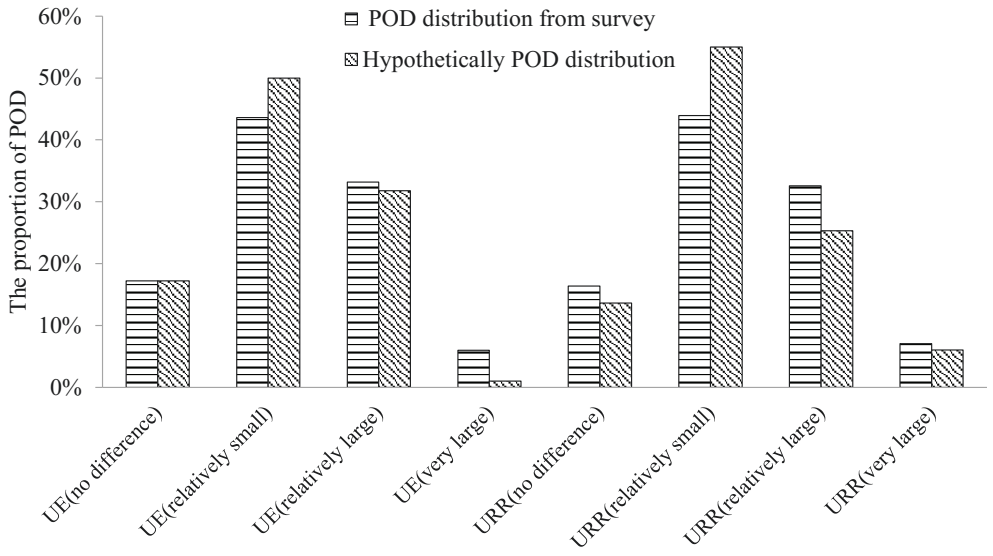


Figure 11: Comparative illustration of the two POD distributions

Figure 12 presents a comparison between the three SRI settings with respect to the two insurance types and different age groups. The figure shows a decrease on SRI for either insurance type and across age groups. This can be explained as follows. Now that with the POD distribution change, many consumers have more trust on the care quality of LWH, they would be more willing to visit LWHs under SRI_REC_2 than SRI_IE. Thus there would be no need to provide SRIs at the same level as previously. Neverthless, since we perform the G2P-SD optimization, the design of SRIs still basically follows the ranking of the *funding efficiency* values.

After comparing the total social cost under the three SRI settings (SRI_IE vs. SRI_REC vs. SRI_REC_2), we find the total social cost under the SRI_REC_2 setting is reduced by 31.98% and 18.54%, respectively, compared with the SRI_IE setting and SRI_REC setting (i.e., 7.30×10^6 under the SRI_IE setting vs. 6.09×10^6 under the SRI_REC vs. 4.93×10^6 under the SRI_REC_2). Further, Figure 13 presents a comparison of the average wait times under the three SRI settings (SRI_IE vs. SRI_REC vs. SRI_REC_2). The figure suggests the possibility of greatly reducing the average HWH wait time, and meanwhile, slightly increasing the average LWH wait time, through more direct ways of varying the POD distribution in the population. Of course, to make the tradeoff between the financial incentivation and some promotional campaign for correcting the quality perception on LWH, it is important to assess the cost of the campaign, and subsequentially, analyze the incremental cost-effectiveness between the above two distinct patient flow alteration mechanisms.

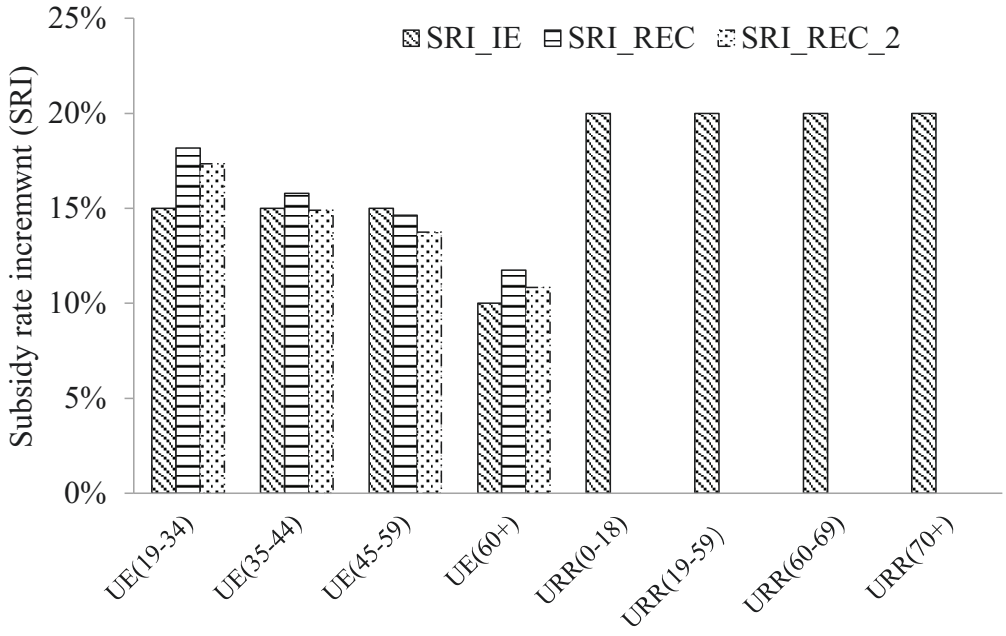


Figure 12: SRI comparison with respect to age group for UE and URR insurance types

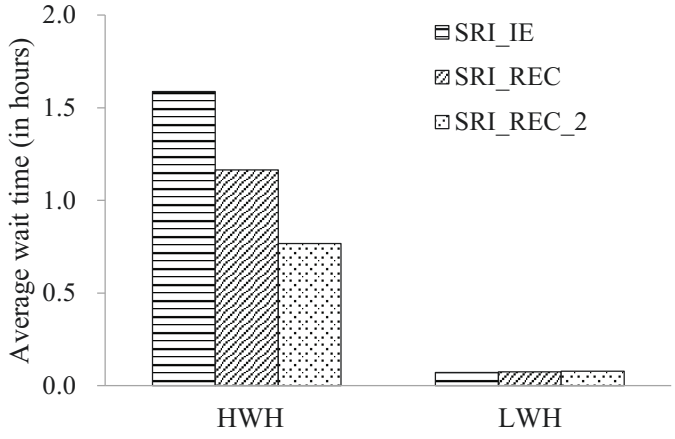


Figure 13: Comparison of the average wait times

Overall, before we give further explanation for our numerical studies, we define that under a certain SRI, patients with a higher probability of choosing LWH is called patients who are more willing to visit LWH under that SRI. And patients who has a greater increase in the probability of choosing LWH brought about by a small increment on a certain SRI is called patients with more sensitivity to that SRI. To give more evidence for the optimal results, we first offer the following conclusions, which are verified by the mathematical explanations in Appendix E. We find the current SRI, the willingness to choose LWH under the current SRI, and the sensitivity to the current SRI, all will have an impact on the optimal G2P-SD design. (1) If the current SRI is the same, patients with higher willingness to go to LWH at the current SRI and less sensitive to the current SRI should be offered a smaller increment on the current SRI. (2) If the current SRI is not the same, patients with larger current SRI, and higher willingness to go to LWH at the current SRI and less sensitivity to the current SRI should be offered a smaller increment on the current SRI. These conclusions can support the interpretation of our G2P-SD optimal design in Study 1 and Study 2.1 – Study 2.3. Taking Study 1 as an example, we give the following explanation. Compared to patients with URR-insurance, UE-insurance patients have lower current SRI, and are more sensitive to their current SRI and less willing to visit LWH, thus they should be offered a larger increment on their current SRI. This is explained that the optimal increment on the current SRI for UE-type insurance are higher than that for URR-type insurance. Second, for UE-insurance patients aged 19–59, the current SRI for them is the same, and older patients are more willing to visit LWH under the same current SRI and are less sensitive to the increment on the current SRI. Therefore, for patients aged 19-34, 35-44, and 45-59, the older the patients are, the smaller the increment is on their current SRIs in the optimal policy, which is consistent with the results in Table 4. Thus, for these three age groups, the older the patients are, the smaller the increment is on their SRIs. Different from them, patients aged 60+, their willingness to visit LWH is the highest, but their current SRI and sensitivity are both smaller, so there is a trade-off between these three factors, and it is necessary to compare the *funding efficiency* of patients aged 60+ with that of patients in other three age groups to determine how their increment on the current SRI should be adjusted. According to the values in Table 4, their *funding efficiency* value is only second to that of people aged 19-34, so the increment on SRI for them is ranked the second. Similarly, the reason for changes on the SRI for four age-specific classes of URR-insurance type is similar to that for UE-insurance type. And the interpretations of the results in Study 2.1 – 2.3 are also similar to the above analysis in Study 1.

5. Conclusions and Future Research

Government subsidization of patients' medical expenses is a must-needed financial incentive mechanism to regulate population-level care access in the healthcare system to maintain good system-level service performance. The hospital system in many countries is hierarchical that

contains diverse hospitals with substantially different amounts of workload. Thus, well-designed subsidy differentials among these hospitals can positively alter each patient's choice on hospital visit so as to rebalance the workload in the multi-hospital system visited by a diverse population of patients. In summary, from the government's perspective, G2P-SD can be a useful lever to induce patients to visit the "right" hospitals.

In this paper, we study the problem of optimal G2P-SD policy design with focus on two-tier hospital system. Our analytic approach contains four parts: (1) developing a binary choice model to characterize the hospital visit behaviors of a heterogeneous patient population; (2) modeling the two-tier system with a large-scale two-level queuing network where the arrivals are specified by the choice model; (3) analyzing the performance measures of the queuing network considering patient balking behavior; and (4) formulating a funding allocation optimization model with the cost of weighted wait time and the government spending as objectives and subject to the minimum workload requirement at each hospital. To the general OR/MS audience, our work presents contributions in three aspects: (1) integration of choice model into optimal subsidy differential pricing; (2) application of queueing model evaluation to healthcare service operations management; and (3) real-world case studies for a country with imminent need.

We administer an online survey and conduct a choice experiment to parameterize the choice models. Our choice model results suggest that several objective attributes such as age and income are influential to patient hospital visit choice. The results also suggest preconceived outcome difference between the two hospital tiers is an important factor. We design case studies based on the realistic situation in Shanghai where we have developed solid partnership with a variety of hospitals along the hospital hierarchy. Through our case study, we verify that (1) optimal design of age-specific G2P-SD policy with our approach can significantly improve the service quality centric system performance; (2) policy tailoring with respect to influential attributes identified in choice model, such as income level, can further improve the system performance; and (3) the notion of *funding efficiency*can serve as a useful indicator to SRI design and budget reallocation. Specifically about the real context of Shanghai, our results suggest that (1) the current SRI, the willingness to choose LWH under the current SRI, and the sensitivity to the current SRI, all will influence the optimal G2P-SD design; (2) it is beneficial to widen the SRIs for the UE-type insurance and narrow the SRIs for the URR-type insurance; and (3) it may be worth investigating the trade-off between financial incentive mechanisms with alternative non-financial mechanisms such as promotional campaigns.

Future research can be pursued along several directions. First, one extension is to introduce additional heterogeneity among hospitals such as distance and and quality level, and further incorporate social welfare concerns in the formulation as consequences of balking. Second, we plan to incorporate some factors such as the estimated waiting time into the utility of balking and to estimate the balking behavior in a discrete choice experiment. Third, the possible extension is to

consider the choice model in which individual-specific variables (such as income) are interacted with subsidy rate in subsidy policy design. Fourth, the possible extension could be to analyze randomness (e.g., time-varying arrival and load-dependent service) more realistically that exists widely in healthcare service systems. Moreover, we plan to extend the dimension of our decision making to include facility location and capacity expansion decisions. Finally to make our work more implementable and practically meaningful, we plan to conduct more appealing choice experiments and develop more comprehensive choice models.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [Grant No. 71432006; 71871138; 71471114] and CEIBS Healthcare Research Fund.

References

[1] L. Hu, Research about the relationship between community hospitals service quality and citizens intention to seek medical care with an example based on Hangzhou, Master's Thesis, School Manage., Zhejiang University, Hangzhou, China.

[2] P. C. Sprivulis, J.-A. Da Silva, I. G. Jacobs, G. A. Jelinek, A. R. Frazer, The association between hospital overcrowding and mortality among patients admitted via western Australian emergency departments, Medical Journal of Australia 184 (5) (2006) 208–212.

[3] R. W. Derlet, J. R. Richards, Overcrowding in the nation's emergency departments: Complex causes and disturbing effects, Annals of Emergency Medicine 35 (1) (2000) 63–68.

[4] C. Visser, G. Marincowitz, I. Govender, G. Ogunbanjo, Reasons for and perceptions of patients with minor ailments bypassing local primary health care facilities, South African Family Practice 57 (6) (2015) 333–336.

[5] Y. Kadooka, A. Asai, A. Enzo, T. Okita, Misuse of emergent healthcare in contemporary Japan, BMC Emergency Medicine 17 (1) (2017) 23.

[6] Q. Qu, P. Guo, R. Lindsey, Comparison of subsidy schemes for reducing waiting times in healthcare systems, Production & Operations Management 26 (11) (2017) 2033–2049.

[7] E. M. Cepolina, A. Farina, Urban car sharing: An overview of relocation strategies, WIT Transactions on the Built Environment 128 (2012) 419–431.

[8] A. Singla, M. Santoni, G. Bartók, P. Mukerji, M. Meenen, A. Krause, Incentivizing users for balancing bike sharing systems., in: AAAI, 2015, pp. 723–729.

[9] S. Aflaki, D. A. Andritsos, Competition and the operational performance of hospitals: The role of hospital objectives, Production & Operations Management 24 (11) (2016) 1812–1832.

[10] W. Chen, Z. G. Zhang, Z. Hua, Analysis of two-tier public service systmes under a government subsidy policy, Computers & Industrial Engineering 90 (2015) 146–157.

[11] S. Çelik, C. Maglaras, Dynamic pricing and lead-time quotation for a multiclass make-to-order queue, Management Science 54 (6) (2008) 1132–1146.

[12] Q. Kong, C.-Y. Lee, C.-P. Teo, Z. Zheng, Scheduling arrivals to a stochastic service delivery system using copositive cones, Operations Research 61 (3) (2013) 711–726.

[13] A. Kuiper, B. Kemper, M. Mandjes, A computational approach to optimized appointment scheduling, Queneing Syetems 79 (1) (2015) 5–36.

[14] T. Cayirli, K. K. Yang, A universal appointment rule with patient classification for service times, no-shows, and walk-ins, Service Science 6 (4) (2014) 274–295.

[15] R. R. Chen, L. W. Robinson, Sequencing and scheduling appointments with potential call-in patients, Production & Operations Management 23 (9) (2014) 1522–1538.

[16] S. Wang, N. Liu, G. Wan, Managing appointment-based services in the presence of walk-in customers, available at SSRN: `http://dx.doi.org/10.2139/ssrn.3104045` (2018).

[17] S. G. Johansen, S. Stidham, Control of arrivals to a stochastic input–output system, Advances in Applied Probability 12 (4) (1980) 972–999.

[18] S. Stidham, Scheduling, routing, and flow control in stochastic networks, in: Stochastic Differential Systems, Stochastic Control Theory and Applications, Springer, 1988, pp. 529–561.

[19] U. Yildirim, J. J. Hasenbein, Admission control and pricing in a queue with batch arrivals, Operations Research Letters 38 (5) (2010) 427–431.

[20] M. E. Ben-Akiva, S. R. Lerman, S. R. Lerman, Discrete Choice Analysis: Theory and Application to Travel Demand, Vol. 9, MIT press, 1985.

[21] S. Mahajan, G. Van Ryzin, Inventory competition under dynamic consumer choice, Operations Research 49 (5) (2001) 646–657.

[22] P. K. Chintagunta, H. S. Nair, Structural workshop paper—discrete-choice models of consumer demand in marketing, Marketing Science 30 (6) (2011) 977–996.

[23] C. K. Anderson, X. Xie, A choice-based dynamic programming approach for setting opaque prices, Production & Operations Management 21 (3) (2012) 590–605.

[24] G. Van Ryzin, G. Vulcano, Simulation-based optimization of virtual nesting controls for network revenue management, Operations Research 56 (4) (2008) 865–880.

[25] R. P. Bagozzi, Y. Yi, Specification, evaluation, and interpretation of structural equation models, Journal of the Academy of Marketing Science 40 (1) (2012) 8–34.

[26] N. Z. Abidin, M. Mamat, B. Dangerfield, J. H. Zulkepli, M. A. Baten, A. Wibowo, Combating obesity through healthy eating behavior: a call for system dynamics optimization, PloS one 9 (12) (2014) e114135.

[27] D. Kahneman, A. Tversky, Prospect theory: An analysis of decision under risk, in: Handbook of the Fundamentals of Financial Decision Making: Part I, World Scientific, 2013, pp. 99–127.

[28] D. A. Hensher, J. M. Rose, W. H. Greene, Applied Choice Analysis: a Primer, Cambridge University Press, 2005.

[29] M. Feehan, M. Walsh, J. Godin, D. Sundwall, M. A. Munger, Patient preferences for healthcare delivery through community pharmacy settings in the USA: A discrete choice study, Journal of Clinical Pharmacy and Therapeutics 42 (6) (2017) 738–749.

[30] J. A. Whitty, E. Kendall, A. Sav, F. Kelly, S. S. McMillan, M. A. King, A. J. Wheeler, Preferences for the delivery of community pharmacy services to help manage chronic conditions, Research in Social and Administrative Pharmacy 11 (2) (2015) 197–215.

[31] A. Scott, C. Bond, J. Inch, A. Grant, Preferences of community pharmacists for extended roles in primary care, Pharmacoeconomics 25 (9) (2007) 783–792.

[32] K. A. Grindrod, C. A. Marra, L. Colley, R. T. Tsuyuki, L. D. Lynd, Pharmacists' preferences for providing patient-centered services: A discrete choice experiment to guide health policy, Annals of Pharmacotherapy 44 (10) (2010) 1554–1564.

[33] A. D. Sinaiko, How do quality information and cost affect patient choice of provider in a tiered network setting? Results from a survey, Health Services Research 46 (2) (2011) 437–456.

[34] K. M. Harris, How do patients choose physicians? Evidence from a national survey of enrollees in employment-related health plans, Health Services Research 38 (2) (2003) 711–732.

[35] J. M. Bronstein, M. A. Morrisey, Bypassing rural hospitals for obstetrics care, Journal of Health Politics, Policy and Law 16 (1) (1991) 87–118.

[36] A. Coulter, N. Le Maistre, L. Henderson, Evaluation of London Patient Choice: Patients' Experience of Choosing Where to Undergo Surgical Treatment, Picker Institute, 2005.

[37] B. J. Borah, A mixed logit model of health care provider choice: Analysis of NSS data for rural India, Health Economics 15 (9) (2006) 915–932.

[38] W.-T. C. Tai, F. W. Porell, E. K. Adams, Hospital choice of rural medicare beneficiaries: patient, hospital attributes, and the patient–physician relationship, Health Services Research 39 (6p1) (2004) 1903–1922.

[39] M. Wensing, H. P. Jung, J. Mainz, F. Olesen, R. Grol, A systematic review of the literature on patient priorities for general practice care. part 1: Description of the research domain, Social science & medicine 47 (10) (1998) 1573–1588.

[40] N. Liu, S. R. Finkelstein, M. E. Kruk, D. Rosenthal, When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling, Management Science 64 (5) (2017) 1975–1996.

[41] A. R. Hole, Modelling heterogeneity in patients' preferences for the attributes of a general practitioner appointment, Journal of health economics 27 (4) (2008) 1078–1094.

[42] S. Cheraghi-Sohi, A. R. Hole, N. Mead, R. McDonald, D. Whalley, P. Bower, M. Roland, What patients want from primary care consultations: a discrete choice experiment to identify patients' priorities, The Annals of Family Medicine 6 (2) (2008) 107–115.

[43] G. Rubin, A. Bate, A. George, P. Shackley, N. Hall, Preferences for access to the gp: a discrete choice experiment, Br J Gen Pract 56 (531) (2006) 743–748.

[44] N. Osadchiy, D. Kc, Are patients patient? the role of time to appointment in patient flow, Production and Operations Management 26 (3) (2017) 469–490.

[45] A. Scott, M. S. Watson, S. Ross, Eliciting preferences of the community for out of hours care provided by general practitioners: a stated preference discrete choice experiment, Social science & medicine 56 (4) (2003) 803–814.

[46] G. Wan, Q. Wang, Two-tier healthcare service systems and cost of waiting for patients, Applied Stochastic Models in Business Industry 33 (2) (2017) 167–183.

[47] V. Denoyel, L. Alfandari, A. Thiele, Optimizing healthcare network design under reference pricing and parameter uncertainty, European Journal of Operational Research 263 (3).

[48] L. Liu, V. G. Kulkarni, Balking and reneging in m/g/s systems exact analysis and approximations, Probability in the Engineering Informational Sciences 22 (3) (2008) 355–371.

[49] P. Guo, C. S. Tang, Y. Wang, M. Zhao, The impact of reimbursement policy on social welfare, revisit rate and waiting time in a public healthcare system: fee-for-service vs. bundled payment, Revisit Rate and Waiting Time in a Public Healthcare System: Fee-for-Service vs. Bundled Payment (October 25, 2017).

[50] Y. Zhang, D. Atkins, Medical facility network design: User-choice and system-optimal models, European Journal of Operational Research 273 (1) (2019) 305–319.

Appendices

Appendix A. Detailed derviations of the queueing network performance measures

With the total arrival rate of an HWH i equipped with capaity c_i^H is $\lambda_i^H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)$, and that of an LWH j equipped with capaity c_i^L is $\lambda_j^L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)$,

$$\lambda_i^H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L) = \frac{c_i^H \lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)}{\sum_{i=1}^{N_H} c_i^H} + p_{LH} \frac{c_i^H \sum_{j=1}^{N_L} \left[\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L) \cdot \left(1 - p_j^{B_L}\right) \cdot c_j^L \Big/ \sum_{j=1}^{N_L} c_j^L \right]}{\sum_{i=1}^{N_H} c_i^H} \tag{19}$$

$$\lambda_j^L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L) = \lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L) \cdot c_j^L \Big/ \sum_{j=1}^{N_L} c_j^L \tag{20}$$

we can derive the utilization rate of an HWH i as

$$\rho_i^H(\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) = \frac{\lambda_i^H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)}{c_i^H \mu_H} \tag{21}$$

$$= \frac{\lambda_H(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)}{\mu_H \sum_{i=1}^{N_H} c_i^H} + p_{LH} \frac{\sum_{j=1}^{N_L} \left[\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L) \cdot \left(1 - p_j^{B_L}\right) \cdot c_j^L \Big/ \sum_{j=1}^{N_L} c_j^L \right]}{\mu_H \sum_{i=1}^{N_H} c_i^H}; \tag{22}$$

$$\rho_j^L(\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)) = \frac{\lambda_j^L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)}{c_j^L \mu_L} = \frac{\lambda_L(\boldsymbol{\sigma}, \mathbf{s}_H, \mathbf{s}_L)}{\mu_L \sum_{j=1}^{N_L} c_j^L}. \tag{23}$$

With the exterior overall arrival rate of class r GC patients at LWHs, i.e., $\lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r)$, the efficient arrival rate of GC patients at an LWH j is obtained as $\lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r)(1-p_j^{B_L})c_j^L / \sum_{j=1}^{N_L} c_j^L$. For a node in equilibrium, its arrival rate and service throughput are equal, the overall throughput of class r patients in all LWHs is

$$T_{m,L}^r(\lambda_L^r(\sigma_r, s_H^r, s_L^r)) = \sum_{j=1}^{N_L} \left[\frac{\lambda_{m,L}^r(\sigma_r, s_H^r, s_L^r)\left(1 - p_i^{B_H}\right) c_j^L}{\sum_{j=1}^{N_L} c_j^L} \right]. \tag{24}$$

Similarly, with the exterior overall arrival rate of class r SC patients at HWHs, i.e., $\lambda_s^r = K_s p_r$, the efficient arrival rate of class r SC patients at an HWH i is $K_s p_r (1-p_j^{B_L}) c_i^H / \sum_{i=1}^{N_H} c_i^H$. And with the exterior overall arrival rate of class r GC patients at HWHs, i.e., $\lambda_{m,H}^r(\sigma_r, s_H^r, s_L^r)$, and the overall throughput of class r patients at LWHs, i.e., $T_{m,L}^r(\lambda_L^r(\sigma_r, s_H^r, s_L^r))$, the efficient arrival rate of GC patients at an HWH i is $(\lambda_{m,H}^r(\sigma_r, s_H^r, s_L^r) + T_{m,L}^r(\lambda_L^r(\sigma_r, s_H^r, s_L^r))\, p_{LH})(1-p_i^{B_H}) c_i^H / \sum_{i=1}^{N_H} c_i^H$.

Thus the overall throughputs of class r GC patients and that of class r SC patients at HWHs are

$$T^r_{m,H}\left(\lambda^r_{m,H}(\sigma_r, s^r_H, s^r_L)\right) = \sum_{i=1}^{N_H}\left[\frac{\left(\lambda^r_{m,H}\left(\sigma_r, s^r_H, s^r_L\right) + T^r_{m,L}\left(\lambda^r_L(\sigma_r, s^r_H, s^r_L)\right)p_{LH}\right)(1-p_i^{B_H})c_i^H}{\sum_{i=1}^{N_H} c_i^H}\right], \tag{25}$$

$$T^r_s = \sum_{i=1}^{N_H}\left[\frac{K_s p_r (1-p_i^{B_H})c_i^H}{\sum_{i=1}^{N_H} c_i^H}\right]. \tag{26}$$

Appendix B. The hospital visit choice behavior questionnaire

In this appendix, we provide the questionnaire we used to survey a cohort of online respondents and model hospital visit choice behavior in Shanghai. The original questionnaire is written in Chinese. We provide its English translation here.

Introduction

You are being invited to take part in a research study about hospital visit choices. Please note that there are no right or wrong answers to any questions in this questionnaire. We are only interested in your opinions and feedback. Your kind and valid response will help the Shanghai municipal government devise better medical reimbursement policy and will help make you feel more satisfied about coverage of your medical care expenses in the future. This questionnaire should take approximately 3-5 minutes to complete.

We assure you that the responses provided by you will not be linked to any personal identifiable information. Your participation in this study is voluntary and you are free to withdraw at any time without penalty. We thank you again for your willingness to participate in this study. Please feel free to contact us if you need any additional information about this project.

1) What is the type of your medical insurance?

   a) Urban Employee (UE) Medical Insurance

   b) Urban and Rural Resident (URR) Medical Insurance

   c) I don't know what kind of insurance I have

   d) No insurance

Section 1: Basic information

The first section of the questionnaire includes questions about your demographics and other related information. We will only use your responses to these questions to compare across survey participants. We assure you that your privacy is protected.

2) What is your gender?

a) Female b) Male

3) Which of the following categories does your age falls into?

a) 0-18 b) 19-34 c) 35-44 d) 60-69 e) over 70

4) What is the highest level of education you have obtained?

a) Primary school or below b) Junior high school

c) High school or vocational school d) Junior college

e) Bachelor degree or above

5) What's your occupation?

a) Unemployed b) Student

c) Employees of state-owned enterprises and institutions d) Self-employed or private owners

e) Employees of private or foreign companies f) Peasant

g) Worker h) Retired

i) Other (Please specify)

6) Which of the following income groups includes your monthly individual income

a) less than 3000 RMB b) 3000-5000 RMB c) 5000-10000 RMB

d) 10000-30000 RMB e) over 30000 RMB

Section 2: Choice scenario

In the following, we will present a scenario where we would like you to choose whether to go to a nearby community-based hospital you could desire imaginatively. Please note that there are no correct or incorrect responses, and your choice should be based on your own preferences, experiences, and specific needs.

Suppose you or your child had fever and cough with headache, muscle pain and other symptoms, you would go to a hospital in need of basic medical service, e.g., an outpatient consultation. Imagine you have two options, either going to a top-rated hospital[1], i.e., an HWH, or a nearby community-based hospital, i.e., an LWH. Please note that healthcare professionals do not think that there is any significant difference in care outcomes between HWH and LWH in the treatment of diseases with the above symptoms.

[1]In China, these hospitals are given a 3A designation.

7) In your opinion, what is the difference in care outcomes between HWH and LWH for outpatient care of mild fever or coughing with headache?

a) no difference b) relatively small

c) quite large d) huge

8) Which of the following categories does your child's age fall into[2]?

a) 0-18 b) over 18 c) no children

9) Assuming that **your child** experienced fever and cough with headache, muscle pain and other symptoms. S/he would thus be in need for going to a children's hospital for basic medical service. Suppose your child is covered by medical insurance[3]. A portion of the total expenses is covered by the government (The proportion paid by the government is called subsidy rate), and the rest is paid by you. The subsidy rate for the HWH- and LWH-tier hospitals are given in the following Table 7. When you make your choice, please take a moment to think about the care outcomes difference between the two hospitals preconceived by you.

Table 7: Would you choose HWH or LWH for your child?

	Alternatives	
	HWH	LWH
Subsidy rate	50 % [4]	$X\%$ [5]
Which one would you choose?		

10) Assuming that **you** experienced fever and cough with headache, muscle pain and other symptoms, and thus would be in need for going to a hospital for basic medical care. Your insurance type is T[6]. A portion of the total expenses is covered by the government (The proportion paid by the government is called subsidy rate), and the rest is paid by you. The subsidy rate for the HWH- and LWH-tier hospitals are given in the following Table 8. When you make your choice, please take a moment to think about the care outcomes difference between the two hospitals preconceived by you.

Appendix C. Interaction effects between patient-specific variables and SRI

[2]The question that follows will depend on the respondent's answer here. The online questionnaire will present question 9 to the respondent if the answer is A; otherwise present question 10.

[3]The insurance type for the minors is always URR.

[4]Currently in Shanghai, the G2P subsidy rate of HWH is always 50% for the minors

[5]The subsidy rate at the LWH is randomly assigned to the respondent within some prespecified range.

[6]T is replaced in real time in the survey according to the respondent's actual insurance type input earlier in the questionnaire.

[7]The corresponding input of the HWH G2P subsidy rate for class r

Table 8: Would you choose HWH or LWH?

	Alternatives	
	HWH	LWH
Subsidy rate	s_r^{H7}	$Y\%^5$
Which one would you choose?		

Table 9: Coefficient estimation results for interaction effects between patient-specific variables and SRI

Insurance Type	**Variable Name**	**Parameter Estimate**	**Standard Error**	***p* –Value**
UE	SRI	13.764	2.484	0.000
	SRI × Age_{35 − 44}	1.412	0.785	0.072
	SRI × Age_{45 − 59}	2.781	0.890	0.002
	SRI × Age_{60+}	3.890	1.422	0.006
	SRI × Income_{$3K - 5K$}	-0.864	2.252	0.704
	SRI × Income_{$5K - 10K$}	-4.152	2.199	0.059
	SRI × Income_{$10K - 30K$}	-6.056	2.269	0.008
	SRI × Income_{$30K+$}	-9.672	3.955	0.014
	SRI × POD_{relatively small}	-2.766	1.046	0.008
	SRI × POD_{relatively big}	-8.639	1.116	0.000
	SRI × POD_{very big}	-8.867	1.615	0.000
	ASC	-1.566	0.154	0.000
URR	SRI	3.873	0.761	0.000
	SRI × Age_{19 − 59}	3.206	0.419	0.000
	SRI × Age_{60 − 69}	4.886	1.208	0.000
	SRI × Age_{70+}	9.739	1.586	0.000
	SRI × Income_{$3K - 5K$}	-0.530	0.593	0.371
	SRI × Income_{$5K - 10K$}	-0.833	0.575	0.147
	SRI × Income_{$10K - 30K$}	-1.087	0.710	0.126
	SRI × Income_{$30K+$}	-3.991	1.664	0.016
	SRI × POD_{relatively small}	-1.140	0.551	0.039
	SRI × POD_{relatively big}	-5.496	0.598	0.000
	SRI × POD_{very big}	-8.639	1.040	0.000
	ASC	-0.559	0.091	0.000

To test the interaction effects between patient-specific variables and SRI, we provided our binary choice model estimation results in Table 9. We discuss in the following how patient-specific variables interact with SRI. We found that for the UE insurance type, age, income, and POD significantly interact with SRI. In particular, when other patient-specific variables being equal, older patients experience higher utility gain when SRI increases, and patients with higher income experience less utility gain when SRI increases. Additionally, when other patient-specific variables being equal, patients whose POD is large experience less utility gain when SRI increases. For the URR insurance type, the interaction effects of SRI with age and POD are significant. Similarly, when other patient-specific variables being equal, older patients experience higher utility gain when SRI increases, and patients whose POD is large experience less utility gain when SRI increases.

Appendix D. Additional information on generation of the hypothetical patient population in the Shanghai-based case study

According to survey on a sample 1% Shanghai population in 2015, we can get the corresponding proportion of different ages of urban and rural populations shown in Table 10.

Table 10: Age distribution of a large-size sample of Shanghai residents covered by either type of insurance

	UE insurance				**URR insurance**			
Age	**19–34**	**35–44**	**45–59**	**60+**	**0–18**	**19–59**	**60–69**	**70+**
Proportion	42.0%	16.4%	22.1%	19.5%	10.5%	66.9%	12.3%	10.3%

The distribution of preconceived outcome difference and income level of the urban and rural populations are from the data of sample statistic in the choice experiment shown in Table 11.

Table 11: Distributions of preconceived outcome difference and income based on the participants from the self-conducted choice experiment

	POD			**Income**	
	UE	**URR**		**UE**	**URR**
No difference	17.2%	16.4%	**<3k**	3.0%	17.8%
Relatively small	43.6%	44.0%	**3k-5k**	26.6%	30.1%
Relatively large	33.2%	32.6%	**5k-10k**	50.7%	38.3%
Very large	6.0%	7.1%	**10k-30k**	19.0%	12.3%
			>30k	0.7%	1.6%

Appendix E. Mathematical explanations about the conclusions on optimal SRI design

Without loss of generality, we assume that the probability of choosing an LWH when a patient is offered an SRI is expressed as $p_L = a_{SRI} * SRI + b_{SRI}$, in which a_{SRI} represents the sensitivity of patients to SRI, and b_{SRI} is a constant related to patient classes and the SRI.

We assume that the total number of patients in the system is n, the subsidy rate at HWH is s_H and the per-capita service payment for patients is C. We also assume that in the current policy, the subsidy rate increment is SRI. Thus the number of patients at HWH denoted by n_H,

the number of patients at LWH denoted by n_L and the total government spending denoted by S_T are respectively expressed in equation (27) – (29).

$$n_H = n * (1 - p_L), \tag{27}$$

$$n_L = n * p_L, \tag{28}$$

$$\begin{aligned} S_T &= n_H * s_H + n_L * (s_H + SRI) \\ &= n * s_H + n * p_L * SRI \end{aligned} \tag{29}$$

If in a new subsidy policy, there is an increment x on the SRI, then the probability of choosing an LWH is changed to $p'_L = a_{SRI} * (SRI + x) + b_{SRI}$. Accordingly, the number of patients at HWH denoted by n'_H, the number of patients at LWH denoted by n'_L and the total government spending denoted by S'_T are respectively expressed in equation (31) - (32).

$$n'_H = n * (1 - p'_L) \tag{30}$$

$$n'_L = n * p'_L \tag{31}$$

$$\begin{aligned} S'_T &= n'_H * s_H + n'_L * (s_H + SRI + x) \\ &= n * s_H + n * p'_L * SRI + n * p'_L * x \end{aligned} \tag{32}$$

1050 We denote the increment of p'_L relative to p_L by Δp_L expressed by equation (33). In the new subsidy policy, the decrease in the number of patients in HWH denoted by Δn_H is expressed in equation (34), while the increase in the number of patients in LWH denoted by Δn_L is expressed by equation (35), and the. And the increment of the total government spending in the new subsidy policy is denoted by ΔS_T, which is expressed in equation (36).

$$\Delta p_L = p'_L - p_L = a_{SRI} * x. \tag{33}$$

$$\Delta n_H = n_H - n'_H = n * (1 - p'_L) - n * (1 - p_L) = n * a_{SRI} * x. \tag{34}$$

$$\Delta n_L = n'_L - n_L = n * p'_L - n * p_L = n * a_{SRI} * x. \tag{35}$$

$$\begin{aligned} \Delta S_T &= S'_L * - S_L \\ &= n * s_H + n * p'_L * SRI + n * p'_L * x - n * s_H - n * p_L * SRI \\ &= n * a_{SRI} * x * SRI + n * p'_L * x \end{aligned} \tag{36}$$

Therefore, an increase of x in the current SRI will reduce the number of patients in the HWH by Δn_H, and the total cost of the government will increase by ΔS_T. Our goal is to reduce wait times in the system, which means we can use financial incentives to guide more patients to LWHs. Therefore, we can define the *funding efficiency* as $f_{S_T} = \Delta n_H / \Delta S_T$. Obviously, to guide patients to LWHs more effectively, patients with higher *funding efficiency* should be given larger increment on SRI.

$$\begin{aligned} f_{S_T} &= \Delta n_H / \Delta S_T \\ &= n * a_{SRI} * x / (n * a_{SRI} * x * SRI + n * p'_L * x) \\ &= a_{SRI} / (a_{SRI} * SRI + p'_L) \end{aligned} \tag{37}$$

From equation 37, when the sensitivity to SRI is the same (i.e., a_{SRI} is the same), we find that the current SRI, the willingness to choose LWH under the current SRI, and the sensitivity to the current SRI, all will have an impact on the optimal G2P-SD design. (1) if the current SRI is the same, the higher willingness to visit LWH (i.e., p'_L is larger), the lower the *funding efficiency* of patients, while the lower willingness to visit LWH (i.e., p'_L is smaller), the higher the *funding efficiency* of patients. As a result, patients who are less willing to visit LWH should have a smaller increment on SRI than patients who are more willing to visit LWH. If the current SRI is not the same: (2) the lower willingness to visit LWH and the smaller the current SRI, the higher the *funding efficiency* of patients. Therefore, the increment on SRI of this kind of patients should also be greater. In reverse, the higher willingness to visit LWH and the larger the current SRI, the smaller the *funding efficiency* of patients. Thus the increment on SRI of this kind of patients should also be smaller; (3) if the willingness to visit LWH is higher, but the current SRI is smaller, or the willingness to visit LWH is lower, but the current SRI is larger, the patient's *funding efficiency* should be judged according to the actual situation.

When the sensitivity to SRI is not the same, we find that (1) if the current SRI is the same, patients who are more willing to visit LWH at current SRI and less sensitive to the current SRI (i.e., a_{SRI} is smaller) should be offered a smaller increment on the current SRI. In reverse, patients who are less willing to visit LWH at the current SRI and more sensitive to the current SRI should be offered a larger increment on the current SRI. If the current SRI is not the same: (2) if the current SRI is larger, the willingness to visit LWH is higher, and the sensitivity to SRI is lower, then the patient's *funding efficiency* is smaller. Thus the increment on SRI of this kind of patients should also be smaller. In reverse, if the current SRI is smaller, the willingness to visit LWH is lower, and the sensitivity to SRI is higher (i.e., a_{SRI} is larger), then the patient's *funding efficiency* is larger.