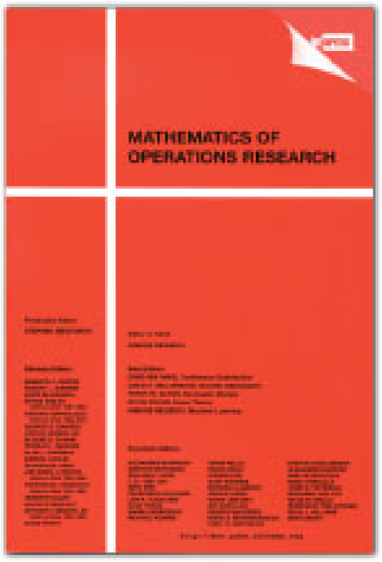

Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimal Provision-After-Wait in Healthcare

Mark Braverman, Jing Chen, Sampath Kannan

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 41, No. 1, February 2016, pp. 352–376
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

http://dx.doi.org/10.1287/moor.2015.0731
© 2016 INFORMS

Optimal Provision-After-Wait in Healthcare

Mark Braverman
Department of Computer Science, Princeton University, Princeton, New Jersey 08544, mbraverm@cs.princeton.edu

Jing Chen
Department of Computer Science, Stony Brook University, Stony Brook, New York 11794, jingchen@cs.stonybrook.edu

Sampath Kannan
Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, kannan@cis.upenn.edu

We investigate computational and mechanism design aspects of allocating medical treatments at hospitals of different costs to patients who each value these hospitals differently. The payer wants to ensure that the total cost of all treatments is at most the budget, B. Access to overdemanded hospitals is rationed through waiting times.

We first show that optimizing social welfare in equilibrium is NP-hard. But if the number of hospitals is small and the budget can be relaxed to $(1+\epsilon)B$ for arbitrarily small ϵ, the optimum under budget B can be achieved efficiently. Next, we show waiting times emerge endogenously from the dynamics between hospitals and patients and the payer doesn't have to explicitly enforce them; all it needs to do is enforce the amount of money paid to each hospital, and the dynamics will converge to the desired waiting times in finite time. Going beyond equilibrium solutions, we investigate the optimization problem over a much larger class of mechanisms. With two hospitals and concave preference profiles of the patients, optimal welfare is actually attained by the randomized assignment, which allocates patients at random and avoids waiting times. Finally, we discuss potential policy implications of our results, followup directions, and open problems.

Keywords: healthcare; budget constraint; waiting times; optimal social welfare; computational complexity; approximation
MSC2000 subject classification: Primary: 90B50; secondary: 68W25, 91B32
OR/MS subject classification: Primary: healthcare; secondary: analysis of algorithms
History: Received February 13, 2014; revised March 2, 2015. Published online in *Articles in Advance* January 5, 2016.

1. Introduction. In this paper, we study computational and mechanism design issues in the context of optimal healthcare provision. Specifically, we consider the setting where waiting times, and not payments, are used to allocate scarce care resources among patients. Waiting times in healthcare provision is an important topic of public debate worldwide. For example, it has a central role in the ongoing debate surrounding the Patient Protection and Affordable Care Act ("Obamacare") in the United States. In a large number of countries with public health coverage financing, including Australia, Canada, Spain, and the United Kingdom, procedures such as elective surgery are rationed by waiting; see Siciliani and Hurst [27] and Gravelle and Siciliani [11]. While in the public perception waiting times are often associated with poor resource management, in the economics literature, it is well understood that queues of consumers will form whenever a good is priced below the good's perceived value, as long as supply is scarce independently of the ultimate distribution mechanism; see Barzel [4], Lindsay and Feigenbaum [21], and Iversen [16]. In particular, waiting times in this context are dictated by economic incentive constraints and not by stochastic fluctuations as in classical queuing theory. Therefore, whenever "correct" monetary pricing is impossible or undesirable, waiting times should be incorporated explicitly into the allocation models.

We focus on providing a single nonurgent healthcare service (such as a particular surgery) to a population of patients, and define the Provision-after-Wait problem for this scenario. In our model, a population of patients arrives in each time unit (say, one month), seeking the desired service at some hospital. There are k hospitals providing the service under different costs. The patients have different preferences about the hospitals, and the composition of the patient population in each time unit is the same. Each patient needs to be served exactly once. The service is fully financed by a third party—a "payer," e.g., the government or an insurer. Therefore the patients' choices of hospitals are not affected by the (monetary) costs. But the payer, taken to be the government in the rest of this paper for concreteness, has a fixed budget B that it is willing to spend on providing the service to the entire patient population in each time unit, and it is unaffordable to let every patient go to his or her favorite hospital (otherwise the provision problem is already solved at the very beginning). Without loss of generality, we assume that the government has enough budget to treat all patients in the cheapest hospital. This can always be achieved by adding a dummy hospital, which has cost 0 and is the least preferred by all patients, representing the option of not getting the service.

The government rations the patients' demand subject to its budget by setting for each hospital H_i a waiting time w_i, measured using the same time unit. Every patient going to H_i has to wait for w_i before being served.

There are no copays, and thus the waiting time is the only cost directly incurred by the patients.[1] We assume that waiting times are known to the patients before they make decisions.[2] Each patient P_j has value v_{ij} for hospital H_i, representing the individual's utility for being treated in H_i right away. Similar to Gravelle and Siciliani [10], we assume that the patients have quasi-linear utilities with respect to waiting time, that is, patient P_j's utility for being treated at H_i with waiting time w_i is $u_{ij} \triangleq v_{ij} - w_i$. The primary reason for this choice is that it is the most natural way to ensure that patients are treated equally by welfare-optimizing mechanisms. Since, as mechanism designers, we do not have full access to the u_{ij}'s of individual patients but can observe waiting times, our welfare loss due to waiting will just be the sum of all the waiting times in the system.[3]

The patients are unrestricted in their choices of hospitals. Thus, at equilibrium, a patient is assigned to a hospital that maximizes his or her utility given the waiting times. The *social welfare* of an equilibrium is defined as the total utility of the patients in each time unit. The government's goal in solving the Provision-after-Wait problem is to find the optimal equilibrium waiting times and assignments of patients to hospitals that maximize social welfare, subject to the budget constraint.

Our model is formally defined in §2. Below, we emphasize three main features of it.

Two noninterchangeable "currencies." Firstly, as money is still involved, the setting leads to two noninterchangeable "currencies" of money and waiting time. This complicates the design problem, conceptually and computationally. We shall see from the first part of our main results, even though money and waiting time are kept separate and only the latter affects the demand, the fact that they cannot be "traded" for each other (thus reducing the setting to one currency) makes the problem much more difficult.

Indirect control of waiting times. Secondly, although waiting time is modeled as a parameter whose optimal value is decided by the government, there is no need for the government to enforce it explicitly. Instead, as we shall show in the second part of our main results, the government can simply decide the amount of money it is willing to pay to each hospital in each time unit, and the desired waiting times at different hospitals will emerge endogenously among the hospitals and the patients. The role of waiting time in our model is similar to that of price in markets. In a market, the price ultimately drives consumers to different purchases, but the producers do not get to dictate it. They can only control the price indirectly by adjusting their supply levels, and the "correct" price will emerge endogenously from the market. This analogy makes it more reasonable to adopt our model in reality: it is more natural for the government to control the amount of money it pays and tell a hospital "I'll only pay you $5,000 each month for this service," than for it to control waiting times and tell a hospital "you have to make each patient using this service wait for three months."

Welfare-burning effect of waiting times. Finally, unlike monetary transfers, nobody benefits from one's waiting time, and thus waiting times represent a net loss in welfare. That is why in our model, the social welfare is defined as the total *utility* of the patients—that is, total value minus total waiting time—differently from auctions, where social welfare is the total *value* of the buyers. The welfare-burning phenomenon is common in the study of resource allocation with waiting times, and is similar to the money-burning mechanisms by Hartline and Roughgarden [14], subject to the important caveat that, in our model, time and money are used and time burnt is not interchangeable with money.

Given the general welfare-burning effect of waiting times, it is very natural to ask whether they can be avoided or reduced via a different allocation mechanism altogether. If monetary payments are not allowed, and patients are free to choose their hospitals, then the (deterministic) equilibrium solution of the Provision-after-Wait problem is the only one possible. What if the government has sufficient control over the patients that it can tell them where to receive their treatment, or can otherwise restrict their options?[4] The simplest such mechanism would be a randomized assignment of patients to available slots, with the probabilities decided by the budget constraint. In such an assignment, we benefit from zero waiting time. On the downside, we incur an efficiency loss: patients may not end up in the hospitals they prefer. How does this randomized assignment mechanism compare to the mechanism where patients are given a free choice and waiting times are used as a rationing

[1] Adding copays to the model would be interesting followup work, but the space of possible models is far vaster with copays. Issues in introducing copays include dealing with different people having different time/money trade-offs, and defining the patients' utility properly (with the usual ethical question: Do people with higher utility for money have lower utility for health, a.k.a. "should poor people count for less"?). In this paper we avoid these problems, since time is fair to everybody and our patients' utility is measured in waiting time equivalents.

[2] For example, the patients can observe the length of the lines before deciding which one to join, or they can be informed explicitly when trying to make an appointment.

[3] We can relax this assumption to allow utility functions of the form $u_{ij} = v_{ij} - U(w_i)$, where $U(w)$ is a function (common to all patients) that maps waiting time w to utility loss caused by waiting w time units.

[4] Possible "soft" mechanisms for doing this are discussed below.

tool? The answer to this question depends on the preference profiles of the patients. Informally speaking, if patients have strong *and diverse* preferences on where to be treated, then the free-choice equilibrium mechanism is better, since efficiency gains due to better allocation offset the inefficiency caused by waiting. At the other extreme, if all patients have similar preferences, then no efficiencies are to be gained from patients' choice, and randomized assignment mechanisms are superior. We further investigate this question in the case of two hospitals, in the third part of our main results.

1.1. Main results.

Finding optimal equilibrium waiting times and assignments. We first study the computational issues in our model, assuming that the government is fully informed about the hospitals' costs and the patients' valuations. The following theorem shows that the Provision-after-Wait problem is hard to solve in general.

Theorem 1. *Finding optimal equilibrium waiting times and assignments is NP-hard.*

The hardness result motivates one to ask whether one can efficiently approximate the welfare of the optimal solution. Interestingly, we show that if we relax the budget constraint to $(1+\epsilon)B$ with an arbitrarily small constant ϵ, we can achieve at least as much welfare as the best B-budget equilibrium solution, using an algorithm whose running time depends on $(\log m)^k$, where m is the number of patients in one time unit and k is, as already mentioned, the number of hospitals.

Theorem 2 (Rephrased). *There is an algorithm that runs in time $O((\log_{1+\epsilon} m)^k \cdot m^4)$ and outputs an equilibrium solution such that the total cost is at most $(1+\epsilon)B$ and the social welfare is at least as high as that of the optimal equilibrium solution with budget B.*

It remains an interesting open problem whether there is a welfare approximation algorithm that does not exceed the budget. Also, it is unknown whether there is an approximation algorithm that is polynomial in k.

Our results are formally presented in §§3 and 4. As will become clear there, our optimization problem resembles the classic knapsack problem in that we try to maximize the total utility subject to a budget constraint, but our problem has an incentive component, which does not exist in knapsack. Also, our problem shares features with classic assignment problems such as unit demand auctions (see, e.g., Demange et al. [7] and Aggarwal et al. [1]), which helps us to derive our approximation result. However, our problem differs from classic assignments in that it is not a priori clear how many patients will be sent to each hospital (or, in the language of auctions, how many items are available for sale).

Letting waiting times emerge endogenously. Next, we show how the desired waiting times and the corresponding optimal social welfare can emerge endogenously as the patients arrive and choose their favorite hospitals. Say the government has decided how to spend its budget for the desired service, by using our approximation algorithm above or other methods. The way of spending the budget can be enforced by setting the *quota* for each hospital, namely, how many patients the government is willing to pay in one time unit (of course, the total quota must be at least the number of patients).

It is natural to assume that the hospitals want to keep waiting times as low as possible, and at time 0, all hospitals have waiting time 0. When the patients arrive along time, they choose which hospital to go according to their own valuations and the current waiting times. If a hospital gets overdemanded, namely, the number of patients going there per time unit exceeds the quota paid by the government, then the line there gets longer and this hospital's waiting time increases accordingly. If the waiting time becomes too high due to previous demand, patients arriving later may choose not to go there and the hospital may become underdemanded, causing its waiting time to decrease. As there may be many waiting time vectors of the hospitals that correspond to equilibrium assignment given the quotas, it is not immediately clear which one the dynamics will converge to (if it converges), and how much social welfare the government can generate from the dynamics.

Assuming the patients' valuations are in a generic position (defined in §5), our following theorem characterizes the structure of the optimal equilibrium given any quotas of the hospitals.

Theorem 3 (Rephrased). *For any quotas of the hospitals, there is a unique optimal equilibrium maximizing social welfare. It has the minimum waiting time vector among all equilibria, and any hospital whose quota is not fully used has waiting time* 0.

Here, we start with the classic result of Shapley and Shubik [26] and Demange et al. [7] on the existence of a unique optimal waiting time vector given the quotas, and arrive at our result via a characterization the patients' demand graph under this waiting time vector.

By Theorem 3, it is reasonable to hope that the optimal equilibrium is the one implemented by the dynamics. Our following theorems show this is indeed the case.

Theorems 4 and 5 (Rephrased). *At any point of time, the waiting time of any hospital will never exceed its waiting time in the optimal equilibrium, and thus the social welfare generated in any time unit will be at least the optimal social welfare given the quotas. The dynamics will always converge to the optimal equilibrium, in time proportional to the number of hospitals, the maximum social welfare of the patients, and the maximum quota of the hospitals.*

Similar to the Hungarian method for classic assignments (see, e.g., Easley and Kleinberg [8]), we analyze the dynamics using a potential function. However, a crucial difference is that in the classic method, the prices never go down (unless all of them are shifted down simultaneously and by the same amount, which does not change the buyers' relative preferences), whereas in our dynamics, the waiting times of different hospitals may change in both directions and in an unsynchronized way.

These results are formally presented in §5.

When is the randomized assignment optimal? Finally, we turn our attention to the enlarged setting where we are not limited to mechanisms that produce equilibrium solutions. The two "extreme" mechanisms are the equilibrium mechanism discussed above that gives the patients free choices, and the randomized assignment mechanism that assigns patients at random to available slots and does not give them any choice. In addition, there are infinitely many *lotteries* in between these extremes. In a lottery, the patients are presented with a set of distributions over hospitals, with an expected waiting time associated with each distribution. Instead of free choices among all possible (distributions of) hospitals, the patients can only choose from the available ones in the lottery, and they make choices to maximize their expected utilities.

Intuitively, if there are no extreme variations among the patients' preferences, the randomized assignment should outperform other mechanisms, since it avoids the deadweight loss of waiting times. We give further evidence suggesting that randomized assignment may be superior in terms of social welfare, by analyzing the case when there are two hospitals.

Let the hospitals be H_0 and H_1 with costs c_0 and c_1, respectively, such that $c_0 < c_1$. We assume without loss of generality that patients going to hospital H_0 face no waiting time.[5] Thus patients who prefer H_0 over H_1 will always choose H_0. We can therefore exclude them from consideration, and focus on patients who prefer H_1 over H_0.

We assume a continuous population of such patients, indexed by the $[0, 1]$ interval. Each patient x is associated with a value $v(x)$, representing how much time he or she is willing to wait to be treated in H_1 instead of H_0. That is, $v(x)$ is the difference between x's utility for being treated at H_1 immediately and x's utility for being treated at H_0 immediately. We rename the patients so that $v(x)$ is a nondecreasing function on $[0, 1]$. Thus, for example, $v(0.5)$ represents the median time that patients preferring H_1 are willing to wait to be treated there. We prove the following theorem in §6.

Theorem 6 (Rephrased). *If $v(x)$ is concave, then no lottery can generate more social welfare than the randomized assignment.*

Roughly speaking, to prove Theorem 6, we proceed by deriving the patients' utilities in terms of their values and the probability distributions only, so that the waiting times disappear from the analysis.

This result shows that for a broad class of preferences, the randomized assignment is welfare maximizing even when waiting times are an option available to the government. As a special case, this shows that randomized assignment has better welfare than the optimal equilibrium solution. It would be interesting to find an analogous sufficient condition for three or more hospitals.

1.2. Discussion and open problems. In this paper, we consider two separate issues. The first one is how to optimally allocate treatments in equilibrium, when the government faces budget constraints and waiting times are used to ration patients' behavior. The second one is whether it may be beneficial to do away with the equilibrium requirements by limiting available options of the patients.

[5] Indeed, positive waiting time at H_0 will give patients incentives to go to the more expensive hospital H_1, and thus increase the total cost while burning more social welfare.

Equilibrium solutions. While finding the optimal equilibrium solution in the Provision-after-Wait problem is NP-hard, our approximation result suggests that this problem might not be as difficult in practice. In many cases, the number of treatment facilities involved is fairly small, making an exponential running time in k feasible. Moreover, in some cases, the “hospitals” are actually treatment alternatives that vary in costs (e.g., physiotherapy is cheaper than knee replacement), in which case k may be as low as 2. For the general case where k can be big, it would be interesting to explore restrictions on the patients’ valuations that would make the exact optimization efficient, such as when the valuations are highly correlated, so that the valuation matrix (v_{ij}) has low rank. There are many questions one can ask about the general complexity of the Provision-after-Wait problem, for example, whether it is strongly NP-hard, whether it has a fully polynomial time approximation scheme (FPTAS), whether it is fixed-parameter tractable in the number of hospitals, etc.

Furthermore, as already mentioned, the connection between our optimization problem with unit demand auctions leads to our approximation result. One might also be able to use this connection to answer other questions about the Provision-after-Wait problem. For example, whether in dynamic settings the system will remain in the patient-optimal equilibrium as the population’s preferences slowly shift over time, whether it is possible to approximate optimal welfare in equilibrium if the government only knows the approximate distribution of patient types in the population, or whether one can design mechanisms such that the patients have incentives to truthfully reveal their valuations and the government does not need to know these valuations to begin with. The last question can also be asked about hospitals: namely, whether the government can elicit the hospitals’ true costs via some mechanisms. This question is particularly interesting given the existence of *rents* in healthcare (see Newhouse [23] for discussion about rents). That is, current prices of certain medical services are substantially higher than providers’ true costs, and thus providers are collecting rents from the government (and other insurers). Rents exist because of the government’s incapability in learning the true costs of medical services and because of its need to meet the reservation prices of providers. Finding true costs and paying hospitals accordingly would thus be helpful in reducing the government’s expenses.

The study of waiting times as a rationing mechanism is closely related to the study of ordeal mechanisms by Alatas et al. [2], where other tools (e.g., excessive bureaucracy) are used in place of waiting times to reduce demand to the supply level.[6] These may be used in settings where queues are not an option, such as school choice. Developing computational mechanism design tools for these settings is a very interesting direction of study.

Beyond equilibria. Our third result looks beyond equilibrium solutions. We give evidence that equilibrium solutions are, in fact, dominated in many cases. One immediate implication is that giving the government power to restrict choice may, in fact, improve overall welfare. Although this is perhaps not surprising, choice restriction may be very difficult or politically infeasible to implement in practice, because patients have an inherent preference for choice, as pointed out by Rosén et al. [24].

There are important indirect ways, however, in which the government may influence choice. One of them is through release (or nonrelease) of quality of care information about providers. The topic of quality of care information is important in theory and in practice. In the United States, for example, Medicare has started to publicly release hospital performance information as part of its pay-for-performance push; see Kahn et al. [17]. The effect that performance reporting has on *provider* incentives has been the subject of much study and discussion; see, e.g., Rosenthal et al. [25], Lindenauer et al. [20], and Gravelle and Sivey [13]. It has even been suggested by Ma and Mak [22] that it would be possible to manipulate reported quality metrics in a way that would force the provider to exert first best quality and cost effort. To the best of our knowledge, there has been no work on the effect of quality reporting on *patient* behaviors.[7]

Inasmuch as quality information influences patients’ choices, it may actually cause harm in the context of allocation using waiting times. Consider a scenario where there are two hospitals, a good one H_1 and a bad one H_0. All patients prefer the good hospital over the bad by the same amount, but they have no a priori knowledge about which is which. As a result, both hospitals will receive half the patients, and waiting time will be zero. If the government reveals that H_1 is the good hospital through its quality of care disclosure, then all patients will prefer H_1 over H_0 by the same amount Δ. Unless H_1 has enough slots for everybody, the waiting time there will have to be Δ, which completely burns social welfare and makes all patients worse off

[6] Note that, in medicine, not all ordeals are necessarily dead weight loss. For example, the famous (and highly demanded) Shouldice hernia clinic in Ontario, Canada requires its patients to lose weight before being admitted for a surgery; see Heskett [15]. Most clinics do not place such a requirement.

[7] da Graça and Masson [5] show that, in special market structures, the consumers may benefit from their uncertainty about the product valuation. But the model is very different.

than when they were ignorant. In effect, before the quality disclosure, uninformed patients implemented the randomized assignment—through free choice. Once the quality information was disclosed, the game moved to the equilibrium solution.

Our results and the discussion above suggest that, in some cases, a population of more informed patients will experience higher waiting times and lower overall utility than uninformed patients. This suggests an unfortunate potential side effect of information disclosure in cases where allocation is done by waiting times. Such a side effect deserves further study since, at the moment, quality information release is regarded as an absolute good. Understanding the optimal structure of information released to the patients in terms of overall welfare (as well as provider-side incentives) is an important and interesting direction of study.

Money-burning mechanisms. In our third result, since there are only two hospitals, each patient's valuation is described by one number and we are considering a *single parameter* setting. With discrete patients, the capacity of the more expensive hospital H_1 is exactly $\lambda = B/c_1$, and the game becomes a unit demand auction with λ copies of the same item.

In this context, Hartline and Roughgarden [14] aim to maximize the same social welfare as ours using money-burning mechanisms, and their results apply to our two-hospital case under their settings. But their results have a very different flavor from ours: the performance of various prior-free money-burning mechanisms considered by Hartline and Roughgarden [14] is analyzed relative to a benchmark $\mathcal{G}$ arising from the collection of i.i.d. distributions of valuations, and $\mathcal{G}$ does not look at the properties of the given valuation profile, such as its concavity.

However, for any fixed valuation profile, a λ-unit p-lottery defined by Hartline and Roughgarden [14] is equivalent to a lottery in our sense (and the randomized assignment is equivalent to the λ-lottery defined there). Thus our result implies that, when the valuation profile is concave, the randomized assignment achieves the best expected social welfare among all λ-unit p-lotteries. This, combined with Corollary 3.6 of Hartline and Roughgarden [14], further implies that for any concave valuation profile, the expected social welfare of the randomized assignment is at least $\mathcal{G}/2$. The relative performance of the randomized assignment and the Random Sampling Optimal Lottery mechanism defined by Hartline and Roughgarden [14] remains unclear, because the latter does not necessarily induce a lottery in our sense, and it is only known that it $O(1)$-approximates $\mathcal{G}$. As commented by Hartline and Roughgarden [14], proving an approximation factor less than 10, say, for their mechanism requires a different approach from the current one. It would be interesting to know, for concave valuation profiles, whether the approximation factor of their mechanism can be improved and how the performances of the two mechanisms compare with each other.

1.3. Additional related work.

The role of waiting time in healthcare can be studied from either the supply side, namely, how waiting times interact with the hospitals' incentives, or the demand side, namely, how they interact with the patients' incentives. Siciliani and Hurst [27] give a thorough analysis of existing policies on reducing waiting times by affecting the incentives of either side. Our model focuses on the demand side, and below, we discuss some other works that also focus on this side.

Gravelle and Siciliani [11] study quality and waiting times with the existence of ex post moral hazard. They assume that the patients are ex ante identical, and that the treatment has *objective* quality levels with which the valuations and the costs are monotonically increasing. But notice that if the patients are identical, rationing by waiting times is bounded to burn a lot of social welfare since at equilibrium every patient has to be treated in the same way—as elaborated in our results. In our model, the patients' valuations can be arbitrarily associated with different hospitals, reflecting *subjective* views they may have, and the hospitals' costs can also be arbitrary and do not necessarily reflect their real quality.

Gravelle and Siciliani [10, 12] also study the effect of waiting time prioritization on social welfare. They consider a single waiting list (or in our language, a single hospital), and the patients are prioritized and may face different waiting times in the same list. In our model, different hospitals may have different waiting times, but we do not discriminate the patients, and at the same hospital, everybody faces the same waiting time. Dawson et al. [6] give experimental evidence on the effect of expanding patient choice of providers on waiting times. In their theoretical model, there are two hospitals and the patients can freely go to the one with shorter waiting time. Thus the patients do not have subjective preferences over hospitals, and waiting time is the only parameter affecting their choices. Moreover, Felder [9] studies the relationship between waiting times and coinsurance, with a single hospital and a single representative consumer.

Leshno [19] studies resource allocation in a domain different from healthcare, where the consumers wait for the stochastic arrival of the items. In contrast to our model and the models discussed above, in this work, waiting time does not burn social welfare, as the total waiting time of the consumers is always the time for enough items to arrive. There are two different types of items to be allocated, and also two types of consumers, respectively, preferring one type of items. A consumer can decide whether to take the arriving item or to continue waiting for the preferred type. The social welfare of the system is measured by the probability that a consumer is matched to the preferred type. Although this is a very different model from ours, it is worth mentioning that the author provides a truthful queuing policy, which is optimal. As we discuss in §1.2, it would be interesting to design a truthful mechanism in our model from which the government can elicit the patients' valuations.

Finally, none of the works mentioned above considers the constraint on the budget of the insurance/resource provider as a parameter affecting waiting times and social welfare.

2. The Provision-after-Wait problem.

Now, we define our model formally. The Provision-after-Wait problem studies how to provide a single healthcare service to a population of patients arriving in each time unit, and is specified by the following parameters:

- The set of *hospitals* is $\{H_1, \ldots, H_k\}$.
- For each $i \in [k]$, the *cost* of H_i per patient is $c_i \in \mathbb{Z}^+$, where $\mathbb{Z}^+$ is the set of nonnegative integers.
- The *number of patients* arriving in each time unit is m.
- The distribution of arriving patients does not change over time, and we denote the *set of patients* arriving in each time unit by $\{P_1, \ldots, P_m\}$.
- For each $i \in [k]$ and $j \in [m]$, the *value* of patient P_j for hospital H_i is $v_{ij} \in \mathbb{Z}^+$.
- An *assignment* of the patients to the hospitals is a triple (w, h, λ), where $w = (w_1, \ldots, w_k) \in (\mathbb{Z}^+)^k$ is the *waiting time vector* of the hospitals, h: $[m] \to [k]$ is the *assignment function*, and $\lambda = (\lambda_1, \ldots, \lambda_k) \in \{1, \ldots, m\}^k$ with $\sum_{i\in[k]} \lambda_i = m$ is the *quota vector*, such that $|h^{-1}(i)| = \lambda_i$ for each $i \in [k]$.

According to such an assignment, patient P_j will receive the service at hospital $H_{h(j)}$ after waiting time $w_{h(j)}$.

- A patient P_j's *utility* under assignment (w, h, λ) is $u_j(w, h, \lambda) \triangleq v_{h(j)j} - w_{h(j)}$, that is, quasi-linear in the waiting time.

The *social welfare* of this assignment is $SW(w, h, \lambda) \triangleq \sum_{j\in[m]} u_j(w, h, \lambda)$.

- The government has *budget* $B \in \mathbb{Z}^+$ per time unit, and an assignment (w, h, λ) is *feasible* if $\sum_{i\in[k]} \lambda_i \cdot c_i \le B$.

For the problem to be interesting, we assume that $mc_{\min} \le B < mc_{\max}$, where $c_{\min}$ and $c_{\max}$ are, respectively, the minimum and the maximum cost of the hospitals.

Remark 1. The hospitals' costs, the patients' valuations, and the waiting times are assumed to be integers without loss of generality. As long as they have finite description, we can always choose proper units so that all of them are integers.

Remark 2. The quota vector of an assignment can be inferred from the assignment function, and thus is redundant. We define it explicitly to ease the discussion of our results.

Since in reality, the government may not be able or willing to force a patient to go to the assigned hospital, it must ensure that wherever it wants that patient to go is indeed the best hospital for the individual, given the waiting times. Accordingly, we have the following definition.

Definition 1. Assignment (w, h, λ) is an *equilibrium assignment* if: (1) it is feasible, (2) for each $j \in [m]$ we have $u_j(w, h, \lambda) \ge 0$, and (3) for each $j \in [m]$ and $i \in [k]$, we have

$$u_j(w, h, \lambda) \ge v_{ij} - w_i.$$

Assignment (w, h, λ) is an *optimal equilibrium assignment* if (1) it is an equilibrium assignment and (2) for any other equilibrium assignment (λ', w', h'),

$$SW(w, h, \lambda) \ge SW(w', h', \lambda').$$

The social welfare of optimal equilibrium assignments is denoted by SW_{OEA}.

We would like to emphasize that, in the healthcare literature, waiting time is recognized as a tool to ration supply by driving down demand. As such, the waiting times at equilibrium *do not* depend on the congestion at the hospitals, but rather on the budget and the patients' "willingness to wait." It is possible that at equilibrium, the number of patients going to a hospital per time unit is smaller than its real capacity,[8] and yet waiting time

[8] That is, the maximum number of patients it is able to handle in one time unit. It is easy to introduce real capacities as additional parameters into our model, and explicitly require that a hospital's quota in an assignment does not exceed its real capacity. But doing so does not make the problem any more interesting: the optimization problem is even harder, and all our results remain true. Thus we simply assume that the real capacities are large enough.

there is nonzero. This is because any shorter waiting time will result in more patients demanding that hospital than allowed by the payer, causing its waiting time to increase. This is demonstrated by the following example.

Assume there are two hospitals, H_0 and H_1, with costs \$500 and \$3,000, respectively.[9] There are the same three types of patients arriving in each month, valuing H_1 for 5, 3, 2, respectively, and all valuing H_0 for 0. The government has budget \$6,000 per month. Both hospitals are capable of handling all three patients immediately. However, if the government lets H_1 be saturated and sends all three patients there, the total cost will be \$9,000, which is unaffordable. It is clear that the government can afford only one patient per month at H_1. Thus at the optimal equilibrium, the waiting time at H_1 must be 3, and only the patient who is willing to wait for 5 will actually be served there. Notice that this patient has to wait even though there is no congestion at all, because of the budget constraint. Notice also that, once the government sets the quota of H_1 to 1, there is no need to enforce the waiting time, since it will automatically increase to 3 as patients arrive. Indeed, with both hospitals' waiting times starting at 0, the first patient arriving in month one will be served at H_1 immediately, the second will wait for one month and be served at H_1 in month two, the third will wait for two months and be served at H_1 in month three (or the patient may go to H_2 immediately if the value of H_1 is 2, but the example will not be too different in that case); and so on. When the waiting time at H_1 is smaller than 3, at least two newly arriving patients (with values 5 and 3) will want to wait there, causing its waiting time to increase by 1. This phenomenon is better explained under continuous time and patient population than discrete cases, and we will formally model and analyze it in §5.

As we are interested in the (existence and) computation of optimal equilibrium assignments, we assume that the government has precise knowledge about the cost of each hospital. We may also assume that the government knows each patient's valuation for each hospital, but we do not need it. In fact, it is enough for the government to know the "distribution" of the k-dimensional valuation vectors of the patients, namely, the fraction of the patients having each particular valuation vector. (How to obtain such information is an interesting mechanism design as well as learning problem.) Once it computes w in the optimal solution, the assignment function h will be automatically implemented by the patients going to their favorite hospitals,[10] and the government need not know where each patient is going.

Notice that it is not enough for the government to know the distribution of the valuations for each single hospital, since the correlations between patients' valuations for different hospitals will affect the optimal outcome. For example, say there are two hospitals H_1 and H_2 with costs $B-1$ and 1, respectively, $(B \gg 1)$, and two patients P_1 and P_2. The valuation vector $(v_{11}, v_{21}, v_{12}, v_{22})$ is either $(10, 0, 4, 6)$ or $(10, 6, 4, 0)$. For each single hospital, the distribution of valuations is the same in the two cases. However, in the former case, the optimal waiting time vector is $(0, 0)$ whereas in the latter it's $(4, 0)$. Thus the optimal solution can't be computed given only the valuation distributions of individual hospitals.

3. The computational complexity of optimal equilibrium assignments. We begin with two easy observations about our model, as a warm-up.

The first observation is that, if the patients have unanimous preferences, namely, $v_{ij} = v_{ij'}$ for each $i \in [k]$ and each $j, j' \in [m]$, then no equilibrium assignment can improve the social welfare of the following trivial one: order the hospitals according to the patients' valuations decreasingly, find the first hospital H_i such that $mc_i \leq B$, and assign all patients to H_i with $w_i = 0$ and $w_{i'} = \max_{i'' \in [k]} v_{i''1}$ for any $i' \neq i$. Indeed, for any equilibrium assignment (w, h, λ), we have $v_{h(j)j} - w_{h(j)} = v_{h(j')j} - w_{h(j')}$ for each $j, j' \in [m]$. Letting $i^* = \arg\min_{i:\, h^{-1}(i) \neq \varnothing} c_i$, λ' be such that $\lambda'_{i^*} = m$ and $\lambda'_i = 0$ for all other i, h' be such that $h'(j) = i^*$ for all j, we have that (w, h', λ') is another equilibrium assignment with the same social welfare as (w, h, λ). Thus it suffices to look for an optimal equilibrium assignment that sends all patients to the same hospital. This is also intuitive: if the patients are all the same, then at equilibrium, the government must make them equally happy, and it can do so by treating them in the same way.

Another observation is that, even if the government only cares about meeting the budget constraint in expectation, is allowed to assign each patient to several hospitals probabilistically (with the total probability summing up to 1), and the patients only care about maximizing their expected utilities, the optimal social welfare in expectation will just be the same as the optimal one obtained by deterministic assignments. This is so because,

[9] In reality, the cheap "hospital" may, in fact, be a cheap service such as a CT scan, while the expensive one may, in fact, be an expensive service such as an MRI. A patient is willing to get either one of them, with different values.

[10] Patients can easily compute which hospitals maximize their utilities, given that they know the hospitals' waiting times and their own valuations. If there is more than one favorite hospital for a patient, we assume that this person goes to the cheapest one, so that the budget constraint is satisfied.

at equilibrium, all the hospitals to which a patient P_j is assigned with positive probability must yield maximum utility for P_j (otherwise, P_j's expected utility can be improved by only going to hospitals that maximize this utility). Thus assigning P_j deterministically to the one with the smallest cost leads to another equilibrium assignment with the same social welfare and still meeting the budget constraint. Accordingly, to maximize social welfare, it suffices to consider only *deterministic assignments*.

The following theorem shows that even the optimal deterministic assignments are hard to find in general.

Theorem 1. *Finding optimal equilibrium assignments is NP-hard.*

Proof. The reduction is from the knapsack problem, which is well known to be *NP*-hard. In this problem, there are k items, $a_1, \ldots, a_k$, and each a_i has value v_i and cost c_i. We are also given a budget B, and the goal is to select a subset of items to maximize their total value, while keeping their total cost less than or equal to B.

We can transform this problem to a Provision-after-Wait problem with $k+1$ hospitals and k patients. Each hospital H_i with $1 \le i \le k$ has cost c_i, and each patient P_i has value v_i for H_i and 0 for all others. Hospital H_{k+1} has cost 0 and is valued 0 by all patients. The government has budget B.

Given an equilibrium assignment (w, h, λ) to the Provision-after-Wait problem, we can construct a solution to the knapsack problem with total value equal to $SW(w, h, \lambda)$—the set $A = \{i\colon h(i) = i\}$ is such a solution. Indeed, without loss of generality, we can assume $h(i) = k+1$ whenever $h(i) \neq i$. By the definition of equilibrium assignments, we can also assume $w_{k+1} = 0$, $w_i = v_i$ if $h(i) = k+1$, and $w_i = 0$ otherwise. Thus $SW(w, h, \lambda) = \sum_{i\in A} v_i$, which is the total value of A in the knapsack problem. As the total cost of (w, h, λ) is $\sum_{i\in A} c_i \le B$, the set A meets the budget constraint in the knapsack problem.

It is easy to see that the other direction is also true, that is, given a solution $A \subseteq [k]$ to the knapsack problem, we can construct an equilibrium assignment (w, h, λ) for the Provision-after-Wait problem whose social welfare equals the total value of A.

Accordingly, an optimal equilibrium assignment to Provision-after-Wait corresponds to an optimal solution to knapsack. □

Remark 3. The NP-hardness of the knapsack problem comes from the need for integrality. Its fractional version can be easily solved using a greedy bang-per-buck approach. But this is not the case in our problem. Indeed, as we have noted, given a fractional equilibrium assignment, we can construct a deterministic one with the same social welfare. Thus for our problem, the fractional version is as hard as the integral version.

Moreover, notice that the knapsack problem is reduced to a very special case of our problem: that is, each patient has positive value for a single hospital and 0 for all others. Thus we believe that the complexity of our problem is not fully captured by knapsack, and that it deserves more investigation in the future. We are tempted to conjecture that our problem is actually strongly NP-hard (and thus does not have an FPTAS), but we do not have a conclusive answer right now.

4. Approximating optimal equilibrium assignments with arbitrarily small deficit.

Although the optimization problem is hard when the numbers of patients and hospitals are large, in practice, we expect the number of hospitals to be small, and it makes sense to solve the problem efficiently in this case.

An easy observation is that optimal equilibrium assignments can be found in time $O(m^k \mathrm{poly}(m, k))$. Indeed, there are at most m^k possible assignment functions $h\colon [m] \to [k]$. For each h and the corresponding quota vector λ satisfying $\sum_{i\in[k]} c_i\lambda_i \le B$, the total value of the patients is fixed, and thus maximizing social welfare is equivalent to minimizing total waiting time. Accordingly, the best equilibrium waiting time vector given h and λ can be found using the linear program below (or one can prove that no feasible waiting time vector exists at equilibrium).

$$\min_{w} \sum_{i\in[k]} w_i\lambda_i$$

$$\text{s.t.}\quad \forall\, j \in [m],\quad i \in [k],\qquad v_{h(j)j} - w_{h(j)} \ge v_{ij} - w_i$$

We then choose h such that the corresponding equilibrium assignment (w, h, λ) maximizes social welfare.

Given the above observation, we are interested in replacing the m^k part with a better bound. As we shall illustrate, if the government is willing to violate its budget constraint by an arbitrarily small fraction, then the problem can be solved much more efficiently.

Definition 2. Let ϵ be a positive constant. An assignment (w, h, λ) is an *equilibrium assignment with ϵ-deficit* if it is an equilibrium assignment with the feasibility condition replaced by the following condition:

$$\sum_{i\in[k]} \lambda_i c_i \le (1+\epsilon)B.$$

We shall construct an algorithm that, in time $O(\log_{1+\epsilon}^k m \cdot (1+\epsilon)^3 m^4)$, finds an equilibrium assignment with ϵ-deficit whose social welfare is at least SW_{OEA}, the social welfare of the optimal equilibrium assignments with budget B. To do so, we first establish a strong connection between the Provision-after-Wait problem and the well-studied problem of unit demand auctions (see, e.g., Demange et al. [7], Aggarwal et al. [1], Ashlagi et al. [3], Easley and Kleinberg [8]).

4.1. A connection between the Provision-after-Wait problem and unit demand auctions. A unit demand auction is specified by n goods (perhaps, including identical ones), m buyers, and the values v_{ij} of each buyer $j \in [m]$ for each good $i \in [n]$. The goal is to find an equilibrium allocation and prices, where buyers get the goods that maximize their utilities given the prices.

If we consider the patients in the Provision-after-Wait problem as buyers who want to buy hospital services using waiting times, our setting looks a lot like a unit demand auction. Except one thing: in our setting the set of goods for sale is unknown. It is natural to consider the k hospitals as k goods, but each one of them has to have a certain number of identical copies, because each hospital may serve more than one patient. One cannot simply model the hospitals as k goods with m copies each, because then the resulted auction will give each patient his or her favorite hospital with zero waiting time, and the budget constraint may be broken.

Notice that, if we were given the quota vector λ in the optimal equilibrium solution of the Provision-after-Wait problem, then we can consider each hospital H_i as λ_i copies of identical goods, and we have a well-defined unit demand auction. Every equilibrium solution to this auction leads to an assignment function h and a waiting time vector w, such that (w, h, λ) is an equilibrium assignment to the original Provision-after-Wait problem. In particular, the budget constraint is satisfied automatically, since we started with a quota vector that meets the budget constraint.

In general, for any quota vector λ such that $\sum_i \lambda_i \geq m$, the problem of finding equilibrium assignments with respect to λ reduces to finding equilibrium prices and allocations in unit demand auctions where each hospital H_i corresponds to λ_i identical goods. If λ meets the budget constraint, namely, $\sum_i c_i \lambda_i \leq B$, then the resulting equilibrium assignment meets the budget constraint.

It is well known that a unit demand auction always has equilibrium prices and allocations, which can be found by the Hungarian method; see Kuhn [18]. The only caution is that, for a hospital to have a well-defined waiting time, the prices of its corresponding goods in the unit demand auction must be all the same. Fortunately, it will become clear in §4.2, at equilibrium identical goods must always have the same price, although this is not explicitly required.

Therefore for each quota vector λ, whether it meets the budget constraint or not, there exists an equilibrium assignment with respect to λ. Following the result of Aggarwal et al. [1], the optimal equilibrium assignment with respect to λ can be computed efficiently, and this will lead to our algorithm for approximating the optimal equilibrium solution of the Provision-after-Wait problem.[11]

4.2. A useful result in multiunit auctions. Our algorithm uses that of Aggarwal et al. [1] for unit demand auctions as a black box, therefore we first recall their result (while using our notation to help establish the connection with our results).

Definition 3. A *unit demand auction*, or simply an *auction* in this paper, is a triple (g, m, v), where the set of goods is $\{1, 2, \ldots, g\}$, the set of bidders is $\{1, 2, \ldots, m\}$, and v is the *valuation matrix*, that is, a $g \times m$ matrix of nonnegative integers. Each v_{ij} denotes the valuation of bidder j for good i.

Given an auction (g, m, v), a *matching* is a triple (u, p, μ), where $u = (u_1, \ldots, u_m) \in (\mathbb{Z}^+)^m$ is the *utility vector*, $p = (p_1, \ldots, p_g) \in (\mathbb{Z}^+)^g$ is the *price vector*, and $\mu \subseteq [g] \times [m]$ is a set of (good, bidder) pairs such that no bidder and no good occur in more than one pair. Bidders and goods that do not appear in any pair in μ are *unmatched*.

Definition 4. Given an auction (g, m, v), a matching (u, p, μ) is *weakly feasible* if for each $(i, j) \in \mu$, we have $u_j = v_{ij} - p_i$, and for each unmatched bidder j, we have $u_j = 0$.

A matching (u, p, μ) is *feasible* if it is weakly feasible and for each unmatched good i, we have $p_i = 0$.

A matching (u, p, μ) is *stable* if for each $(i, j) \in [g] \times [m]$, we have $u_j \geq v_{ij} - p_i$.

A matching (u^*, p^*, μ^*) is *bidder optimal* if (1) it is stable and feasible and (2) for every matching (u, p, μ) that is stable and weakly feasible, and for every bidder j, we have $u_j^* \geq u_j$.

[11] Although equilibrium assignments can be efficiently computed given λ, the problem of deciding the "correct" λ makes the Provision-after-Wait problem hard, even in very special cases, as shown in §3.

Aggarwal et al. [1] construct an algorithm, StableMatch, which, given an auction (g, m, v), outputs a bidder-optimal matching (u^*, p^*, μ^*) in time $O(mg^3)$.

Notice that the original definitions in Aggarwal et al. [1] have for each good-bidder pair a reserve price and a maximum price. In our model, we do not need them, so the definitions above are more succinct than the original ones. In fact, as pointed out by Aggarwal et al. [1], with maximum prices, there may be no bidder-optimal matching. But without them, such a matching always exists, as shown by Demange et al. [7].

Notice also that Aggarwal et al. [1] do not distinguish between weak feasibility and feasibility. But it is easy to see that their algorithm and its analysis still apply under our definitions. We shall use these two notions when analyzing our algorithm.

Next, we establish two properties for the matching (u^*, p^*, μ^*) output by StableMatch.

- *Property* 1. If $g \geq m$, then without loss of generality, we can assume that (u^*, p^*, μ^*) has no unmatched bidder.

Indeed, if there exists an unmatched bidder j, then there must exist an unmatched good i (since $g \geq m$). Since (u^*, p^*, μ^*) is bidder optimal, we have $u^*_j = 0$, $p^*_i = 0$, and $u^*_j \geq v_{ij} - p^*_i$. Thus we have $v_{ij} = 0$, and the matching $(u^*, p^*, \mu^* \cup \{(i, j)\})$ is another bidder-optimal matching.

- *Property* 2. If two goods i, i' are identical, namely, $v_{ij} = v_{i'j}$ for each bidder j, then $p^*_i = p^*_{i'}$.

Indeed, if both goods are unmatched, then $p^*_i = p^*_{i'} = 0$. Otherwise, say, $(i, j) \in \mu^*$. By definition, $u^*_j = v_{ij} - p^*_i \geq v_{i'j} - p^*_{i'}$. As $v_{ij} = v_{i'j}$, we have $p^*_i \leq p^*_{i'}$. If i' is unmatched, then $p^*_{i'} = 0$, implying $p^*_i = 0$. If $(i', j') \in \mu^*$, then similarly we have $p^*_{i'} \leq p^*_i$, and thus $p^*_i = p^*_{i'}$ again.

4.3. Our algorithm for approximating optimal equilibrium assignments.

Now, we are ready to construct our algorithm for approximating optimal equilibrium assignments. The algorithm takes as input the number of patients m, the number of hospitals k, the hospitals' costs $c_1, \ldots, c_k$, the patients' valuations v_{ij}'s for the hospitals, the budget B, and a small constant $\epsilon > 0$. Letting (w, h, λ) be an optimal equilibrium assignment, the algorithm works by guessing λ, constructing a multiunit auction based on the guessed vector, computing the bidder-optimal matching using StableMatch, and extracting the waiting time vector and the assignment function from the matching.

More precisely, let $L \triangleq \lceil \log_{1+\epsilon} m \rceil$, $C_0 \triangleq 0$, and $C_l \triangleq \lfloor (1+\epsilon)^l \rfloor$ for each $l = 1, \ldots, L$. The algorithm examines all the vectors $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_k) \in \{C_0, C_1, \ldots, C_L\}^k$ one by one, say, lexicographically.

If $\sum_{i\in[k]} \hat{\lambda}_i \notin [m, (1+\epsilon)m]$ or if $\sum_{i\in[k]} \hat{\lambda}_i c_i > (1+\epsilon)B$, the algorithm disregards this vector and moves to the next. Otherwise, it constructs an auction $(g, m, \hat{v})$ as follows. The set of patients corresponds to the set of bidders; each hospital H_i corresponds to $\hat{\lambda}_i$ copies of identical goods $H_{i1}, \ldots, H_{i\hat{\lambda}_i}$, thus $g = \sum_{i\in[k]} \hat{\lambda}_i$; the valuation matrix $\hat{v}$ has rows indexed by $\{ir\colon i \in [k], r \in [\hat{\lambda}_i]\}$, columns indexed by $[m]$, and for each $j \in [m]$, $i \in [k]$, and $r \in [\hat{\lambda}_i]$, $\hat{v}_{ir,j} = v_{ij}$.

The algorithm then runs StableMatch with input $(g, m, \hat{v})$ to generate the bidder-optimal matching (u^*, p^*, μ^*), and extracts the waiting time vector $\hat{w}$ and the assignment function $\hat{h}$ as follows. For each hospital H_i, let $\hat{w}_i = p^*_{i1}$. For each patient P_j, let H_{ir} be the unique good to which P_j is matched (by Property 1 in §4.2 such a good always exists) according to μ^*, and let $\hat{h}(j) = i$. The triple $(\hat{w}, \hat{h}, \hat{\lambda})$ may not be an assignment as $\sum_{i\in[k]} \hat{\lambda}_i$ may be larger than m, but there is a unique quota vector $\hat{\lambda}'$ such that $(\hat{w}, \hat{h}, \hat{\lambda}')$ is an assignment.

The algorithm computes the social welfare of the assignment $(\hat{w}, \hat{h}, \hat{\lambda}')$ for each $\hat{\lambda}$ that is not disregarded, and output the assignment (w^*, h^*, λ^*) with the maximum social welfare.

We prove the following theorem.

Theorem 2. *Our algorithm runs in time $O(\log^k_{1+\epsilon} m \cdot m^4)$, and outputs an equilibrium assignment with ϵ-deficit (w^*, h^*, λ^*) such that $SW(w^*, h^*, \lambda^*) \geq SW_{\text{OEA}}$.*

Proof. The running time of the algorithm can be immediately seen. Indeed, if a vector $\hat{\lambda}$ is not disregarded, then it takes $O(mg) = O(m^2)$ time to construct the auction as $g \in [m, (1+\epsilon)m]$, $O(mg^3) = O(m^4)$ time to run StableMatch, and $O(m)$ time to extract the assignment. Accordingly, it takes $O(m^4)$ time to examine a single vector $\hat{\lambda}$, and there are $O(\log^k_{1+\epsilon} m)$ vectors in total.

The remaining part of the theorem follows from the two lemmas below.

Lemma 1. (w^*, h^*, λ^*) *is an equilibrium assignment with* ϵ*-deficit.*

Proof. In fact, we show that for each vector $\hat\lambda$ that is not disregarded, the extracted assignment $(\hat w, \hat h, \hat\lambda')$ is an equilibrium assignment with ϵ-deficit. To see why this is true, first notice that $\sum_{i\in[k]} \hat\lambda_i c_i \le (1+\epsilon)B$ by the construction of the algorithm, thus

$$\sum_{i\in[k]} \hat\lambda'_i c_i \le \sum_{i\in[k]} \hat\lambda_i c_i \le (1+\epsilon)B. \tag{1}$$

Second, for each $j\in[m]$, letting $H_{\hat h(j)r}$ be the good matched to P_j according to μ^*, we have

$$u_j(\hat w, \hat h, \hat\lambda') = v_{\hat h(j)j} - \hat w_{\hat h(j)} = \hat v_{\hat h(j)r,\,j} - p^*_{\hat h(j)1} = \hat v_{\hat h(j)r,\,j} - p^*_{\hat h(j)r} = u^*_j \ge 0, \tag{2}$$

where the third equality is because of Property 2 in §4.2 (in particular, $H_{\hat h(j)1}$ and $H_{\hat h(j)r}$ are identical goods, and $p^*_{\hat h(j)1} = p^*_{\hat h(j)r}$), and the other equalities/inequality are by definition.

Third, since (u^*, p^*, μ^*) is a bidder-optimal matching for auction $(g, m, \hat v)$, we have that for each $j\in[m]$, $i\in[k]$, and $r\in[\hat\lambda_i]$,

$$u^*_j \ge \hat v_{ir,\,j} - p^*_{ir} = v_{ij} - p^*_{i1} = v_{ij} - \hat w_i,$$

and thus

$$u_j(\hat w, \hat h, \hat\lambda') = u^*_j \ge v_{ij} - \hat w_i. \tag{3}$$

Equations (1)–(3) together imply that every $(\hat w, \hat h, \hat\lambda')$ is an equilibrium assignment with ϵ-deficit, and so is (w^*, h^*, λ^*). □

Lemma 2. $SW(w^*, h^*, \lambda^*) \ge SW_{\mathrm{OEA}}$.

Proof. To see why this is true, arbitrarily fix an optimal equilibrium assignment (w, h, λ). Notice that for each hospital H_i, there exists a "good guess" $\hat\lambda_i \in \{C_0, \ldots, C_L\}$ such that

$$\lambda_i \le \hat\lambda_i \le (1+\epsilon)\lambda_i.$$

Since λ satisfies $\sum_{i\in[k]} \lambda_i = m$ and $\sum_{i\in[k]} \lambda_i c_i \le B$, the vector $\hat\lambda = (\hat\lambda_1, \ldots, \hat\lambda_k)$ satisfies

$$\sum_{i\in[k]} \hat\lambda_i \in [m, (1+\epsilon)m] \qquad \text{and} \qquad \sum_{i\in[k]} \hat\lambda_i c_i \le (1+\epsilon)B.$$

Thus it won't be disregarded by the algorithm. Let $(g, m, \hat v)$ be the auction constructed from $\hat\lambda$, (u^*, p^*, μ^*) the output of StableMatch under input $(g, m, \hat v)$, and $(\hat w, \hat h, \hat\lambda')$ the assignment extracted from (u^*, p^*, μ^*). Following the same reasoning shown in Equation (2), we have that for each $j\in[m]$, $u_j(\hat w, \hat h, \hat\lambda') = u^*_j$. Thus

$$SW(\hat w, \hat h, \hat\lambda') = \sum_{j\in[m]} u^*_j. \tag{4}$$

From (w, h, λ), we construct a matching (u, p, μ) for the auction $(g, m, \hat v)$ as follows. For each bidder j, we have $u_j = v_{h(j)j} - w_{h(j)}$; for each good H_{ir} with $i\in[k]$ and $r\in[\hat\lambda_i]$, we have $p_{ir} = w_i$; and for each hospital H_i, letting $j_1 \le j_2 \le \cdots \le j_{\lambda_i}$ be the patients assigned to H_i by h, we have $\mu = \{(j_r, ir)\colon i\in[k], r\in[\lambda_i]\}$.

It is easy to verify that the so constructed (u, p, μ) is stable and weakly feasible, thus by the optimality of u^*, we have that for each $j\in[m]$,

$$u^*_j \ge u_j. \tag{5}$$

Moreover, for the same reason as Equation (4), we have

$$SW(w, h, \lambda) = \sum_{j\in[m]} u_j. \tag{6}$$

Equations (4)–(6) together imply

$$SW(\hat w, \hat h, \hat\lambda') \ge SW(w, h, \lambda) = SW_{\mathrm{OEA}}$$

as we want to show. □

In sum, Theorem 2 holds. □

Remark 4. By running our algorithm with input budget $B/(1+\epsilon)$, we obtain an assignment whose budget is at most B and whose social welfare is at least the optimal social welfare with budget $B/(1+\epsilon)$. However, this social welfare may be much smaller than the optimal social welfare with budget B. That is why we insist on having a deficit instead of meeting the budget constraint strictly.

5. The endogenous emergence of waiting times. Next, we study the dynamics between hospitals and patients. We shall consider continuous changes of waiting times, and in the discussion below, the patients' valuations and the waiting times can be any nonnegative reals, not necessarily integers. We show that, in our model, when the patients' valuations are in some generic position, the only thing the government needs to enforce is the amount of money it is willing to pay to each hospital, which can be equivalently enforced by the quota vector. Given the quotas, the optimal waiting times and the optimal social welfare will emerge endogenously from the dynamics.

5.1. The uniqueness of the optimal equilibrium. We start by defining the generic position of the patients and studying the structure of the optimal equilibrium under it. Following Ashlagi et al. [3], we have the definition below.

Definition 5. The patients $\{P_1, \ldots, P_m\}$ with valuations $(v_{ij})_{i\in[k],\, j\in[m]}$ are *independent* if there do not exist two different subsets S and T of the multiset $\{v_{ij}\colon i \in [k],\ j \in [m]\}$ such that S and T both contain positive numbers and $\sum_{v\in S} v = \sum_{v'\in T} v'$.

Notice that the above definition of independent patients is weaker than the typical definition of generic position, which rules out any relevant equality relation among the valuations. Notice also that it is easy to perturb the numbers in the proof of Theorem 1, so that the resulted Provision-after-Wait problem is generic. Thus the optimization problem is still NP-hard in the generic case. But our results below apply to any λ, which may be obtained via approximation algorithms or heuristics.

Let λ be a quota vector with $\sum_{i\in[k]} \lambda_i \geq m$.[12] Recall that given λ, the Provision-after-Wait problem reduces to a unit demand auction. Thus following Shapley and Shubik [26] and Demange et al. [7], among all equilibrium waiting time vectors with respect to λ, there is a unique one that simultaneously minimizes the waiting time at each hospital and maximizes the utility of each patient.[13] Denoting this minimum waiting time vector by $\bar{w}$, we prove the following theorem.

Theorem 3. *Assuming the patients are independent, there is a unique equilibrium assignment with respect to λ and $\bar{w}$. Moreover, denoting this equilibrium by $(\bar{w}, \bar{h}, \lambda)$, we have that $\min_{i\in[k]} \bar{w}_i = 0$, and that at this equilibrium every hospital with positive waiting time is saturated, namely, $|\bar{h}^{-1}(i)| = \lambda_i$ whenever $\bar{w}_i > 0$.*

Proof. Without loss of generality, we assume $\lambda_i > 0$ for each $i \in [k]$. Consider the demand graph G given $\bar{w}$, that is, a bipartite graph with k nodes on one side for the hospitals and m nodes on the other side for the patients. For each $i \in [k]$ and $j \in [m]$, the edge (i, j) is in G if and only if H_i maximizes P_j's utility, namely, $v_{ij} - \bar{w}_i = \max_{i'\in[k]} v_{i'j} - \bar{w}_{i'}$. By definition, any equilibrium assignment must assign each patient P_j to an adjacent hospital H_i. Thus it suffices to show that within each connected component of G there is only one equilibrium assignment. We start by proving the following claim.

Claim 1. *There is no cycle in G.*

Proof. For the sake of contradiction, assume there exists a (necessarily even length) cycle $(i_1, j_1, i_2, j_2, \ldots, i_l, j_l, i_1)$, where i_r's are hospitals and j_r's are patients. By the construction of G, we have that for each $r \in [l]$, H_{i_r} and $H_{i_{r+1}}$ maximize P_{j_r}'s utility, with $l+1$ defined to be 1. Thus

$$v_{i_r j_r} - \bar{w}_{i_r} = v_{i_{r+1} j_r} - \bar{w}_{i_{r+1}}.$$

Summing all l equations together, we have

$$\sum_{r\in[l]} (v_{i_r j_r} - \bar{w}_{i_r}) = \sum_{r\in[l]} (v_{i_{r+1} j_r} - \bar{w}_{i_{r+1}}),$$

therefore

$$\sum_{r\in[l]} v_{i_r j_r} - \sum_{r\in[l]} \bar{w}_{i_r} = \sum_{r\in[l]} v_{i_{r+1} j_r} - \sum_{r\in[l]} \bar{w}_{i_{r+1}}.$$

[12] Notice that we do not require that λ satisfies the budget constraint, and our results apply to such λs as well.

[13] Notice that this is the waiting time vector computed by the StableMatch algorithm of Aggarwal et al. [1].

As $\sum_{r\in[l]}\bar{w}_{i_r}=\sum_{r\in[l]}\bar{w}_{i_{r+1}}$, we have

$$\sum_{r\in[l]}v_{i_r j_r}=\sum_{r\in[l]}v_{i_{r+1}j_r}.$$

Accordingly, we have found two different subsets $\{v_{i_r j_r}\colon r\in[l]\}$ and $\{v_{i_{r+1}j_r}\colon r\in[l]\}$ that sum up to the same value, contradicting the hypothesis that the patients are independent. □

Following Claim 1, the connected components of G are all trees. Similarly, we have the following:

Claim 2. *Each connected component of G contains at most one hospital with waiting time* 0.

Proof. Again, for the sake of contradiction, assume there is a connected component with two different hospitals H_i and $H_{i'}$ such that $\bar{w}_i=\bar{w}_{i'}=0$. Accordingly, there is a path $(i_1,j_1,i_2,j_2,\ldots,i_l)$, where i_r's are hospitals and j_r's are patients such that $i_1=i$ and $i_l=i'$. Similar to the proof of Claim 1, for each $r<l$, we have

$$v_{i_r j_r}-\bar{w}_{i_r}=v_{i_{r+1}j_r}-\bar{w}_{i_{r+1}}.$$

Summing all $l-1$ equations together, we have

$$\sum_{r=1}^{l-1}v_{i_r j_r}-\sum_{r=1}^{l-1}\bar{w}_{i_r}=\sum_{r=1}^{l-1}v_{i_{r+1}j_r}-\sum_{r=1}^{l-1}\bar{w}_{i_{r+1}}.$$

As $\bar{w}_{i_1}=\bar{w}_{i_l}=0$, the above equation implies

$$\sum_{r=1}^{l-1}v_{i_r j_r}-\sum_{r=2}^{l-1}\bar{w}_{i_r}=\sum_{r=1}^{l-1}v_{i_{r+1}j_r}-\sum_{r=2}^{l-1}\bar{w}_{i_r},$$

and thus

$$\sum_{r=1}^{l-1}v_{i_r j_r}=\sum_{r=1}^{l-1}v_{i_{r+1}j_r},$$

again contradicting the hypothesis that the patients are independent. □

Claim 2 and the following claim together imply that each connected component of G has exactly one hospital with waiting time 0.

Claim 3. *Each connected component of G has at least one hospital with waiting time* 0.

Proof. By contradiction. Assume there is a component C such that $\bar{w}_i>0$ for each H_i in C. Let

$$\epsilon_1=\min_{H_i\in C}\bar{w}_i.$$

Notice that for each P_j not in C, by definition, the best utility that j can get from hospitals in C is strictly less than $u_j^{\max}$, the best utility that j can get from his or her favorite hospital. Let

$$\epsilon_2=\min_{P_j\notin C}\left[u_j^{\max}-\max_{H_i\in C}(v_{ij}-\bar{w}_i)\right].$$

We have $\epsilon_1>0$ and $\epsilon_2>0$. Let $\epsilon=\min\{\epsilon_1,\epsilon_2\}/2$, $w'_i=\bar{w}_i-\epsilon$ for each $H_i\in C$, and $w'=(\bar{w}_{-C},w'_C)$. That is, w' is $\bar{w}$ with all waiting times of hospitals in C reduced by ϵ. As $\epsilon<\epsilon_1$, w' is a valid waiting time vector.

Notice that for any equilibrium assignment $(\bar{w},h,\lambda)$, the assignment (w',h,λ) is still an equilibrium. Indeed, when the waiting time vector changes from $\bar{w}$ to w', for each patient P_j, his or her utility at every hospital $H_i\in C$ increases by ϵ, and his or her utility at every other hospital remains the same. For $P_j\notin C$, $\epsilon<\epsilon_2$, and thus the best utility j gets from C is still smaller than $u_j^{\max}$, which is j's utility at $H_h(j)\notin C$. For $P_j\in C$, we have $H_{h(j)}\in C$ as well, and $H_{h(j)}$ still maximizes j's utility after the increase.

Accordingly, w' is another equilibrium waiting time vector. But $w'_i<\bar{w}_i$ for each $H_i\in C$ and $w'_i=\bar{w}_i$ for each $H_i\notin C$, contradicting the hypothesis that $\bar{w}$ minimizes the waiting time of each hospital among all equilibrium waiting time vectors. Therefore Claim 3 holds. □

Following Claims 1–3, each connected component C can be considered as a tree rooted at the unique hospital with waiting time 0, with hospitals and patients alternating along each path. Based on this structure, we show that there is only one way of assigning the patients to the hospitals at equilibrium in C. To do so, we need the following.

Claim 4. *For each hospital $H_i \in C$ with $\bar{w}_i > 0$, the degree of H_i in G is strictly larger than its quota λ_i.*

The proof is similar to that of Claim 3: if the degree of some $H_i \in C$ is at most λ_i, then we can find a proper value $\epsilon \in (0, \bar{w}_i)$ such that the vector $w' \triangleq (\bar{w}_{-i}, \bar{w}_i - \epsilon)$ is still an equilibrium waiting time vector. Indeed, with properly chosen ϵ, for every equilibrium $(\bar{w}, h, \lambda)$, let h' be the assignment such that $h'(j) = i$ if P_j is adjacent to H_i (this is doable because the degree of H_i is at most λ_i), and $h'(j) = h(j)$ otherwise. Then, (w', h', λ) is another equilibrium. But this contradicts the hypothesis that $\bar{w}$ minimizes the waiting time of each hospital among all equilibrium waiting time vectors. The formal analysis is omitted.

Following Claim 4, we have that the leaves of tree C are all patients. Indeed, if there is a hospital with degree 1 and positive waiting time, then its quota is 0, contradicting our original assumption that all hospitals have positive quotas. Accordingly, at every equilibrium, every patient at a leaf must be assigned to his or her preceding hospital, as this is the only one maximizing the patient's utility. Letting H_i be a nonroot hospital whose descendants are all leaves, we have that the number of descendants of H_i, denoted by d_i, is at most λ_i, otherwise no equilibrium exists. As $\bar{w}_i > 0$, by Claim 4, we have that the degree of H_i is strictly larger than λ_i, which implies $d_i \geq \lambda_i$. Accordingly, H_i uses up all its quota to serve its descendants, and the patient P_j preceding H_i must be assigned to his or her preceding hospital.

Repeating the above reasoning in a bottom-up way along the tree, we have that there is only one way of assigning the patients to hospitals at equilibrium with respect to λ and $\bar{w}$, that is, patients are assigned to their predecessors in G, and hospitals with positive waiting times are saturated by their descendants. Thus Theorem 3 holds. □

By definition, the equilibrium $(\bar{w}, \bar{h}, \lambda)$ maximizes social welfare with respect to λ, thus it is reasonable to assume that this is the equilibrium that the government aims to implement.

5.2. The dynamics between hospitals and patients. We now show that given λ, the waiting time vector $\bar{w}$ will endogenously emerge from the dynamics between hospitals and patients, and so will $\bar{h}$. We consider a continuous-time dynamics, where the patient population arrives continuously and uniformly along time (which is consistent with our discrete model). In such a dynamic, the quota vector λ represents the *service rate* of the hospitals that the government is willing to pay for. Namely, for each hospital H_i, the total number of patients paid by the government in any time interval (t_1, t_2) is at most $\lambda_i(t_2 - t_1)$.[14]

The set of patients in previous sections, $\{P_1, \ldots, P_m\}$ with valuations $(v_{ij})_{i\in[k],\, j\in[m]}$, now represents the set of *types* of the arriving patients. That is, although the patient population goes to infinity, there are only finitely many types of them. Every type has *arrival rate* 1: by any time t, the number of patients that have arrived is mt, where t of them are of type P_1 (i.e., with valuation $(v_{1j}, \ldots, v_{kj})$), and another t of them are of type P_2, etc. We say that the patient population is *independent* if $\{P_1, \ldots, P_m\}$ is independent. Notice that, in general, there may be different P_j and $P_{j'}$ with the same valuation, and the number of patients of a particular type by time t may be larger than t. But when the population is independent, any different P_j and $P_{j'}$ must have different valuations, and indeed represent different types. Below, we consider independent population.

Let $w(t) \triangleq (w_1(t), \ldots, w_k(t))$ be the nonnegative waiting time vector of the hospitals at time t such that $w(0) = (0, \ldots, 0)$. A patient of type P_j arriving at time t chooses a hospital H_i maximizing the individual's utility given $w(t)$, and will be served there at time $t + w_i(t)$.[15] To break ties consistently throughout time, we impose a partial ordering over the hospitals, according to their positions in the demand graph G with respect to $\bar{w}$. In particular, if H_i and $H_{i'}$ are in the same connected component of G and H_i precedes $H_{i'}$, then at any time t and for any patient of type P_j whose utility is maximized at H_i and $H_{i'}$ given $w(t)$, we assume that P_j does not choose H_i. If H_i and $H_{i'}$ are in different connected components, then P_j can choose one

[14] The budget constraint B now represents the spending rate of the government: the total amount of money the government can afford by time t is Bt. But as already said, our conclusion in this section holds even when λ does not satisfy the budget constraint. Thus we shall not talk about the budget constraint in the remaining part of this section.

[15] Therefore the patients are served in a first-in-first-out queue.

arbitrarily, or even split the population of this type arbitrarily between H_i and $H_{i'}$, as indicated by the definition below.

Definition 6. For any i, j, t, the *demand rate of P_j for H_i at time t*, denoted by $d_{ij}(t)$, is a number in $[0, 1]$ such that,

- $\sum_{i\in[k]} d_{ij}(t) = 1$ for all j,
- $d_{ij}(t) > 0$ only if H_i maximizes P_j's utility at time t, and there is no other hospital $H_{i'}$ preceded by H_i in the same connected component of G that does so.

The *demand rate for H_i at time t* is $d_i(t) \triangleq \sum_{j\in[m]} d_{ij}(t)$.

The fractional values of the d_{ij}'s indicate how the patients of the same type will split between all hospitals maximizing their utilities. For example, $d_{ij}(t) = 1/3$ means that fixing the current waiting times, in the long run, a third of the patients of type P_j will choose H_i. Notice that we do not completely specify how the patients should make their decisions when there are ties, and yet our results hold no matter how these ties are broken.

Because the patients arrive continuously under a constant rate, their effect on the waiting times at any point of time is infinitesimal, and $w(t)$ is continuous. By definition, within an arbitrarily small time interval $(t, t+\delta)$, the number of patients choosing H_i is $d_i(t)\delta$. Since the number of patients served by H_i in time δ is $\lambda_i\delta$, the waiting time will not change if $d_i(t) = \lambda_i$ (i.e., if the demand rate matches the service rate), and will change by $(d_i(t)\delta - \lambda_i\delta)/\lambda_i$ otherwise, unless $w_i(t) = 0$ and $d_i(t) < \lambda_i$, in which case $w_i(t+\delta)$ will remain 0. That is,

$$w_i(t+\delta) - w_i(t) = \begin{cases} \left(\dfrac{d_i(t)}{\lambda_i} - 1\right)\delta & \text{if } w_i(t) > 0 \text{ or } d_i(t) \geq \lambda_i, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Accordingly, for each $i \in [k]$ the right derivative of $w_i(t)$ is

$$\frac{d_+ w_i(t)}{dt} = \begin{cases} \lim\limits_{\delta\to 0} \dfrac{w_i(t+\delta) - w_i(t)}{\delta} = \dfrac{d_i(t)}{\lambda_i} - 1 & \text{if } w_i(t) > 0 \text{ or } d_i(t) \geq \lambda_i, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Notice that for particular tie-breaking rules, the function $d_i(t)$ may not be continuous, and thus $w_i(t)$ may not be differentiable. But we can always define its right derivative as above.

We say that $w(t)$ is *at most* $\bar{w}$, written as $w(t) \leq \bar{w}$, if $w_i(t) \leq \bar{w}_i$ for each $i \in [k]$. Moreover, we say that $w(t)$ is *smaller than* $\bar{w}$, written as $w(t) < \bar{w}$, if the above inequality holds for some $i \in [k]$. The following two theorems show that the dynamics will always converge to $\bar{w}$ in finite time, and will never exceed $\bar{w}$ before converging.

Theorem 4. *When the patient population is independent, then*:
(1) *$w(t) \leq \bar{w}$ for any $t \geq 0$*;
(2) *if $w(t) = \bar{w}$, then $d_+ w_i(t)/dt = 0$ for any $i \in [k]$*; *and*
(3) *if $w(t) < \bar{w}$, then there exists $i \in [k]$ such that $d_+ w_i(t)/dt > 0$.*

Proof. To prove Statement (1), it suffices to show the following.

Claim 5. *For any $t \geq 0$ and $i \in [k]$, if $w(t) \leq \bar{w}$ and $w_i(t) = \bar{w}_i$, then $d_i(t) \leq \lambda_i$ and $w_i(t)$ will not increase.*

Proof. Since $|\bar{h}^{-1}(i)| \leq \lambda_i$ by the definition of equilibrium $(\bar{w}, \bar{h}, \lambda)$, it suffices to show

$$d_i(t) \leq |\bar{h}^{-1}(i)|.$$

Since

$$d_i(t) = \sum_{j\in[m]} d_{ij}(t) = \sum_{j:\,\bar{h}(j)=i} d_{ij}(t) + \sum_{j:\,\bar{h}(j)\neq i} d_{ij}(t) \leq \sum_{j:\,\bar{h}(j)=i} 1 + \sum_{j:\,\bar{h}(j)\neq i} d_{ij}(t) = |\bar{h}^{-1}(i)| + \sum_{j:\,\bar{h}(j)\neq i} d_{ij}(t),$$

it suffices to show that for any $j \in [m]$,

$$\text{if} \quad \bar{h}(j) \neq i, \quad \text{then} \quad d_{ij}(t) = 0.$$

To do so, arbitrarily fix a type P_j such that $\bar{h}(j) \neq i$. If $v_{ij} - w_i(t) < \max_{i'}\{v_{i'j} - w_{i'}(t)\}$, then H_i does not maximize P_j's utility under $w(t)$, and thus $d_{ij}(t) = 0$. Assume now

$$v_{ij} - w_i(t) = \max_{i'} \{v_{i'j} - w_{i'}(t)\}.$$

Notice that

$$v_{ij} - w_i(t) = v_{ij} - \bar{w}_i \leq v_{\bar{h}(j)j} - \bar{w}_{\bar{h}(j)} \leq v_{\bar{h}(j)j} - w_{\bar{h}(j)}(t) \leq \max_{i'} \{v_{i'j} - w_{i'}(t)\},$$

where the equality is because $w_i(t) = \bar{w}_i$, the first inequality is because hospital $H_{\bar{h}(j)}$ maximizes P_j's utility under $\bar{w}$, and the second is because $w_{\bar{h}(j)}(t) \leq \bar{w}_{\bar{h}(j)}$ by hypothesis. Thus all the inequalities above are actually equalities, that is,

$$v_{ij} - w_i(t) = v_{ij} - \bar{w}_i = v_{\bar{h}(j)j} - \bar{w}_{\bar{h}(j)} = v_{\bar{h}(j)j} - w_{\bar{h}(j)}(t) = \max_{i'} \{v_{i'j} - w_{i'}(t)\}.$$

On the one hand, the second equality implies that H_i and $H_{\bar{h}(j)}$ are adjacent to P_j in the demand graph G under $\bar{w}$. Since P_j is assigned to $H_{\bar{h}(j)}$, following the analysis of Theorem 3, it must be the case that $H_{\bar{h}(j)}$ precedes P_j and P_j precedes H_i in G. On the other hand, the last equality implies that $H_{\bar{h}(j)}$ also maximizes the utility of P_j given $w(t)$, and thus P_j will not choose H_i according to the tie-breaking rule. Accordingly, $d_{ij}(t) = 0$ as we wanted to show.

In sum, we have $d_i(t) \leq |\bar{h}^{-1}(i)| \leq \lambda_i$, which implies that $w_i(t)$ will not increase by the definition of the dynamics. Therefore Claim 5 holds. □

Since $w(0) = (0, \ldots, 0)$ and $w(t)$ is continuous, Claim 5 implies that $w(t) \leq \bar{w}$ for any $t \geq 0$, and Statement (1) holds.

Statement (2) simply follows from the fact that, when $w(t) = \bar{w}$, the patients choose their hospitals according to the unique equilibrium $(\bar{w}, \bar{h}, \lambda)$, and thus $d_i(t) = |\bar{h}^{-1}(i)| = \lambda_i$ for every i such that $\bar{w}_i > 0$, and $d_i(t) = |\bar{h}^{-1}(i)| \leq \lambda_i$ for every i such that $\bar{w}_i = 0$.

Finally, Statement (3) is equivalent to the following claim, which we prove below.

Claim 6. *If $w(t) < \bar{w}$, then there exists $i \in [k]$ such that $d_i(t) > \lambda_i$.*

Proof. For the sake of contradiction, assume $d_i(t) \leq \lambda_i$ $\forall\, i$. We shall construct a new demand vector $d' = (d'_{ij})_{i\in[k],\, j\in[m]}$ such that

$$d'_{ij} \in \{0, 1\} \quad \forall\, i, j \qquad \text{and} \qquad d'_i \triangleq \sum_j d'_{ij} \leq \lambda_i \quad \forall\, i.$$

Notice that d' induces an equilibrium assignment with waiting time $w(t)$, where each P_j is assigned to the unique hospital H_i with $d'_{ij} = 1$. This contradicts the fact that $\bar{w}$ is the minimum equilibrium waiting time vector with respect to λ.

To find the desired d', consider the demand graph $G(t)$ with respect to $w(t)$. For each H_i and P_j, $d_{ij}(t) > 0$ implies that H_i and P_j are adjacent in $G(t)$. Since the patient population is independent, $G(t)$ is a forest with hospitals and patients alternating along each path, as shown in the proof of Theorem 3.

The construction of d' starts from $G(t)$, processes and removes its nodes step-by-step and in a bottom-up way, and assigns patients to hospitals using a greedy method. More precisely, we initialize

$$d'_{ij} = 0 \quad \forall\, i, j, \qquad d_{ij} = d_{ij}(t) \quad \forall\, i, j, \qquad \text{and} \qquad \lambda'_i = \lambda_i \quad \forall\, i.$$

At any time of the construction, d'_{ij} represents the demand of a processed patient, d_{ij} represents that of a remaining patient,

$$d'_i \triangleq \sum_{j\in[m]} d'_{ij} \quad \forall\, i,$$

and represents the total demand for a hospital from the processed patients,

$$d_i \triangleq \sum_{j:\, P_j \text{ is adjacent to } H_i} d_{ij} \quad \forall\, H_i \text{ in the graph}$$

and represents the total demand for a remaining hospital from the remaining patients, and λ'_i is an integer, which represents H_i's remaining quota after some patients have been assigned to it. It will be invariant that

$$d'_i + \lambda'_i = \lambda_i \quad \forall\, i, \qquad d_i \leq \lambda'_i \quad \forall\, i, \qquad \text{and} \qquad \sum_{i:\, H_i \text{ adjacent to } P_j} d_{ij} = 1 \quad \forall\, P_j \text{ in the graph}. \tag{9}$$

It is easy to see that Equation (9) holds at the beginning.

In each step of the construction, arbitrarily choose a leaf with the longest path from its root in the remaining graph. We distinguish two cases.

Case 1. The chosen leaf is a patient, denoted by P_{j^*}.

This is the simpler case. Letting the unique adjacent hospital be H_{i^*}, by Equation (9), we have

$$d_{ij^*} = 0 \quad \forall\, i \neq i^* \qquad \text{and} \qquad d_{i^*j^*} = 1 \leq d_{i^*} \leq \lambda'_{i^*}.$$

Set $d'_{i^*j^*} = 1$, $d_{i^*j^*} = 0$, and $\lambda'_{i^*} = \lambda'_{i^*} - 1$, and remove P_{j^*} from the graph. That is, P_{j^*} is assigned to H_{i^*} and occupies 1 quota there. Notice that the invariance remains: indeed, d'_{i^*} increases by 1 and λ'_{i^*} decreases by 1, d_{i^*} and λ'_{i^*} decrease by 1, and everything else remains unchanged.

Case 2. The chosen leaf is a hospital, denoted by H_{i^*}.

This is the more complicated case. Letting the unique adjacent patient be P_{j^*}, we have

$$0 \leq d_{i^*j^*} = d_{i^*} \leq \lambda'_{i^*}.$$

On the one hand, if $\lambda'_{i^*} \geq 1$ (that is, H_{i^*} still has quota for one more patient), then set $d'_{i^*j^*} = 1$, $d_{ij^*} = 0\ \forall\, i$, and $\lambda'_{i^*} = \lambda'_{i^*} - 1$. Remove P_{j^*} and its children (which are all leaves, since H_{i^*} has the longest path from the root) from the graph. That is, P_{j^*} is assigned to H_{i^*}, and for any other hospital H_i with P_{j^*} being the only adjacent patient, no patient will be assigned to it any more. Notice that the invariance remains: indeed, d'_{i^*} increases by 1; λ'_{i^*} decreases by 1; $d_{i^*} = d_{i^*j^*} = 0$; λ'_{i^*} is nonnegative; for any $i \neq i^*$, d_i either decreases or remains unchanged; and everything else remains unchanged.

On the other hand, if $\lambda'_{i^*} = 0$, then $d_{i^*j^*} = d_{i^*} = 0$ by Equation (9). That is, no remaining patient wants H_{i^*}. We simply remove H_{i^*} from the graph, keeping the invariance.

Notice that we finish processing all the nodes after at most $m + k$ steps. In the end, all the d'_{ij}'s are either 0 or 1, and $d'_i \leq \lambda_i\ \forall\, i$, as desired. Thus Claim 6 holds. □

Accordingly, Statement (3) holds, and so does Theorem 4. □

Theorem 4 shows that the waiting times in the dynamics will continue increasing before they reach $\bar{w}$, and will stop changing once they reach $\bar{w}$. The only thing remains to show is that, the evolution speed of the dynamics will not go to 0 as time increases, so that it will indeed reach $\bar{w}$ in a finite amount of time. More precisely, letting $MSW = \sum_{j\in[m]} \max_{i\in[k]} v_{ij}$ and $\lambda_{\max} = \max_{i\in[k]} \lambda_i$, we have the following theorem.

Theorem 5. *When the patient population is independent, the dynamics converges to $\bar{w}$ in time at most $2k\lambda_{\max} MSW$.*

Proof. Similar to the Hungarian method (see, e.g., Easley and Kleinberg [8]), we consider the following potential function:

$$P(t) \triangleq \sum_{i\in[k]} \lambda_i w_i(t) + \sum_{j\in[m]} u_j(t),$$

where $u_j(t) \triangleq \max_{i\in[k]}(v_{ij} - w_i(t))$. Since $w_i(t)$ is continuous for each $i \in [k]$, we have that $u_j(t)$ is continuous for each $j \in [m]$ and $P(t)$ is continuous as well.

By Theorem 3, we have $\min_{i\in[k]} \bar{w}_i = 0$. By Theorem 4, we have that, before the dynamics converges, $(0, \ldots, 0) \leq w(t) < \bar{w}$ for any t, and thus $\min_{i\in[k]} w_i(t) = \min_{i\in[k]} \bar{w}_i = 0$. Accordingly, $u_j(t) \geq 0$ for each P_j, and $P(t) \geq 0$.

It is easy to see that $P(0) = MSW$. Thus it suffices to prove that $P(t)$ strictly decreases, and the local decreasing rate is at least $1/(k\lambda_{\max})$.

To do so, notice that

$$\begin{aligned} P(t) &= \sum_i \lambda_i w_i(t) + \sum_j \sum_i d_{ij}(t)(v_{ij} - w_i(t)) \\ &= \sum_i \lambda_i w_i(t) - \sum_i \Bigl(\sum_j d_{ij}(t)\Bigr) w_i(t) + \sum_{i,j} d_{ij}(t) v_{ij} \\ &= \sum_i (\lambda_i - d_i(t)) w_i(t) + \sum_{i,j} d_{ij}(t) v_{ij}. \end{aligned}$$

Thus for arbitrarily small $\delta > 0$, by definition, we have

$$\begin{aligned}
&P(t+\delta) - P(t)\\
&\quad= \sum_i (\lambda_i - d_i(t+\delta))w_i(t+\delta) - (\lambda_i - d_i(t))w_i(t) + \sum_{i,j} d_{ij}(t+\delta)v_{ij} - \sum_{i,j} d_{ij}(t)v_{ij}\\
&\quad= \sum_i (w_i(t+\delta) - w_i(t))(\lambda_i - d_i(t)) - \sum_i w_i(t+\delta)d_i(t+\delta) + \sum_i w_i(t+\delta)d_i(t)\\
&\qquad+ \sum_{i,j} d_{ij}(t+\delta)v_{ij} - \sum_{i,j} d_{ij}(t)v_{ij}\\
&\quad= \sum_i (w_i(t+\delta) - w_i(t))(\lambda_i - d_i(t)) + \sum_{i,j} d_{ij}(t+\delta)v_{ij} - \sum_{i,j} d_{ij}(t+\delta)w_i(t+\delta)\\
&\qquad- \sum_{i,j} d_{ij}(t)v_{ij} + \sum_{i,j} d_{ij}(t)w_i(t+\delta)\\
&\quad= \sum_i (w_i(t+\delta) - w_i(t))(\lambda_i - d_i(t)) + \sum_{i,j} (d_{ij}(t+\delta) - d_{ij}(t))(v_{ij} - w_i(t+\delta)).
\end{aligned}$$

Since $w(t)$ is continuous, $\lim_{\delta\to 0}[v_{ij} - w_i(t+\delta)] = v_{ij} - w_i(t)$. Accordingly, for any i, j such that $v_{ij} - w_i(t) < u_j(t)$, we have $v_{ij} - w_i(t+\delta) < u_j(t)$ for arbitrarily small δ. Since the patients only choose hospitals that maximize their utilities, $d_{ij}(t) = d_{ij}(t+\delta) = 0$. That is, for each P_j,

$$\sum_{i:\, v_{ij} - w_i(t) = u_j(t)} d_{ij}(t) = \sum_{i:\, v_{ij} - w_i(t) = u_j(t)} d_{ij}(t+\delta) = 1.$$

Combining this equation with Equation (7), we have

$$\begin{aligned}
&\lim_{\delta\to 0} \frac{P(t+\delta) - P(t)}{\delta}\\
&\quad= -\sum_{i:\, w_i(t) > 0 \text{ or } d_i(t) \geq \lambda_i} \frac{(d_i(t) - \lambda_i)^2}{\lambda_i} + \sum_j u_j(t) \lim_{\delta\to 0} \frac{\sum_{i:\, v_{ij} - w_i(t) = u_j(t)} (d_{ij}(t+\delta) - d_{ij}(t))}{\delta}\\
&\quad= -\sum_{i:\, w_i(t) > 0 \text{ or } d_i(t) \geq \lambda_i} \frac{(d_i(t) - \lambda_i)^2}{\lambda_i} + \sum_j u_j(t) \lim_{\delta\to 0} \frac{1-1}{\delta}\\
&\quad= -\sum_{i:\, w_i(t) > 0 \text{ or } d_i(t) \geq \lambda_i} \frac{(d_i(t) - \lambda_i)^2}{\lambda_i}. \qquad (10)
\end{aligned}$$

To upper bound the last part of Equation (10), we consider the set of hospitals

$$B \triangleq \left\{ i \colon \bar{w}_i - w_i(t) = \max_{i' \in [k]} \{\bar{w}_{i'} - w_{i'}(t)\} \right\}.$$

As $w(t) < \bar{w}$ before the dynamics converges, there exists i such that $\bar{w}_i - w_i(t) > 0$. Thus for any i with $\bar{w}_i = 0$, $i \notin B$. By Theorem 3, hospitals in B are all saturated under $(\bar{w}, \bar{h}, \lambda)$, that is,

$$|\bar{h}^{-1}(i)| = \lambda_i \quad \forall\, i \in B.$$

For any patient j with $\bar{h}(j) \in B$, we have

$$\sum_{i\in B} d_{ij}(t) = 1,$$

because when the waiting times change from $\bar{w}$ to $w(t)$, the utilities of j at hospitals in B become *strictly* more advantageous against the patient's utilities at hospitals not in B: indeed, by definition, the waiting times for hospitals in B decrease the most from $\bar{w}$ to $w(t)$. Thus

$$\sum_{j:\, \bar{h}(j)\in B} \sum_{i\in B} d_{ij}(t) = \sum_{j:\, \bar{h}(j)\in B} 1 = \sum_{i\in B} |\bar{h}^{-1}(i)| = \sum_{i\in B} \lambda_i.$$

Let BP be the set of patients j such that $\bar{h}(j) \notin B$ but j is adjacent to a hospital in B in the demand graph of $\bar{w}$ (BP for "boundary patients"). Notice that $BP \neq \varnothing$ as $B \neq [k]$. For any $j \in BP$, we again have $\sum_{i\in B} d_{ij}(t) = 1$, for a similar reason as before—that is, at $\bar{w}$ patient j is indifferent between the best hospital (for him or her) in B and the best not in B, and from $\bar{w}$ to $w(t)$, the hospitals in B become strictly more advantageous. Accordingly,

$$\sum_{i\in B} d_i(t) \geq \sum_{j:\bar{h}(j)\in B}\sum_{i\in B} d_{ij}(t) + \sum_{j\in BP}\sum_{i\in B} d_{ij}(t) = \sum_{i\in B}\lambda_i + \sum_{j\in BP} 1 \geq \sum_{i\in B}\lambda_i + 1.$$

Let $B' \triangleq \{i\in B \mid d_i(t)\geq \lambda_i\}$. By definition, we have $\sum_{i\in B\setminus B'}\lambda_i \geq \sum_{i\in B\setminus B'} d_i(t)$, and therefore

$$\sum_{i\in B'} d_i(t) \geq \sum_{i\in B'}\lambda_i + 1.$$

Thus, by the concavity of the function x^2 and Jensen's inequality, we have

$$\begin{aligned}\lim_{\delta\to 0}\frac{P(t+\delta)-P(t)}{\delta} &= -\sum_{i:\,w_i(t)>0 \text{ or } d_i(t)\geq\lambda_i}\frac{(d_i(t)-\lambda_i)^2}{\lambda_i} \leq -\sum_{i:\,d_i(t)\geq\lambda_i}\frac{(d_i(t)-\lambda_i)^2}{\lambda_i}\\ &\leq -\sum_{i\in B'}\frac{(d_i(t)-\lambda_i)^2}{\lambda_i} \leq -\sum_{i\in B'}\frac{(d_i(t)-\lambda_i)^2}{\lambda_{\max}} = -\frac{|B'|}{\lambda_{\max}}\cdot\frac{1}{|B'|}\cdot\sum_{i\in B'}(d_i(t)-\lambda_i)^2\\ &\leq -\frac{|B'|}{\lambda_{\max}}\cdot\left(\frac{\sum_{i\in B'}(d_i(t)-\lambda_i)}{|B'|}\right)^2 \leq -\frac{|B'|}{\lambda_{\max}}\cdot\left(\frac{1}{|B'|}\right)^2 \leq -\frac{1}{k\lambda_{\max}},\end{aligned} \tag{11}$$

for any time t before the dynamics converges.

Finally, letting $T = 2k\lambda_{\max}MSW$ and assuming that the dynamics does not converge before time T, we show that $P(T)=0$, and thus the dynamics must converge at time T. For the sake of contradiction, assume $P(T)>0$. We have

$$P(T)-P(0) > 0 - MSW = -\frac{T}{2k\lambda_{\max}}.$$

Let

$$t^* \triangleq \sup\left\{t\colon\, t\leq T,\ P(t)-P(0)\leq -\frac{t}{2k\lambda_{\max}}\right\}.$$

Since $P(t)$ is continuous, we have

$$P(t^*)-P(0)\leq -\frac{t^*}{2k\lambda_{\max}},$$

and thus $t^*<T$. By our hypothesis, the dynamic does not converge before T, thus by Inequality (11), there exists $\delta\in(0, T-t^*)$ such that

$$P(t^*+\delta)-P(t^*)\leq -\frac{\delta}{2k\lambda_{\max}}.$$

Letting $t' = t^*+\delta$, we have $t^*<t'<T$ and

$$P(t')-P(0) = P(t^*+\delta)-P(t^*)+P(t^*)-P(0)\leq -\frac{t^*+\delta}{2k\lambda_{\max}} = -\frac{t'}{2k\lambda_{\max}},$$

contradicting the definition of t^*. Therefore $P(T)=0$ and the dynamics converges to $\bar{w}$ in time at most T, as desired. □

Remark 5. Although the potential function used in the above proof is similar to that used in the Hungarian method for unit demand auctions, the analysis is different. For example, the potential function in the latter measures the total price paid at each time step, while ours measures the "budgeted" total waiting time $\sum_i \lambda_i w_i(t)$, which can be very different from the total waiting time. Moreover, in the latter, the prices of the goods for sale never go down, making the analysis much easier, while in our dynamics, the waiting times may go up and down, depending on the demands.

6. The optimality of the randomized assignment. Although waiting time is widely used to ration demand in economic settings, it may burn a lot of social welfare, since the time waited is not beneficial to anybody. Therefore, in this section, we study different allocation schemes in healthcare and give evidence that the government can avoid the welfare-burning effect of waiting times by limiting the choices available to the patients. In particular, we show that the randomized assignment is actually optimal in terms of social welfare in many cases.

Following our discussion in §1, we consider the case of two hospitals, a "good" one H_1 and a "bad" one H_0, with costs $c_1 > c_0$. As already said, whoever prefers H_0 can be directly assigned there and we no longer consider them in our setting. The patients preferring H_1 are indexed by the interval $[0, 1]$, and each patient x is associated with a value $v(x)$, indicating how long the individual is willing to wait at H_1 to be treated there instead of H_0. We assume that the patients have been renamed and normalized, so that $v(x)$ is nondecreasing and $v(0) = 0$. Since the number of patients is infinite, we talk about the *cost density* $c_i(x)$ of each hospital, rather than the cost for serving a single patient. Without loss of generality, $c_1(x) \equiv 1$ and $c_0(x) \equiv 0$. The government has budget $B \in (0, 1)$, meaning that at most a B fraction of the patients can be served at H_1. The government's goal is to maximize the expected social welfare subject to the requirement that the budget constraint is satisfied in expectation.

In the randomized assignment, the government assigns each patient to H_1 with probability p and waiting time 0. The budget constraint gives

$$\int_0^1 pc_1(x)\,dx = p = B,$$

and the corresponding social welfare, denoted by SW_r, is

$$SW_r = \int_0^1 pv(x)\,dx = B\int_0^1 v(x)\,dx. \tag{12}$$

Below, we compare this social welfare with that of lotteries.

Definition 7. A *contract* is a pair (p, w), where $p \in [0, 1]$ is the probability of assigning a patient to H_1, and $w \ge 0$ is the waiting time for that patient at H_1.

A *lottery* consists of a set of contracts, denoted by the domain $D \subseteq [0, 1]$ of the probabilities, and the waiting time function $w(p)$ defined over D.

Given a contract $C = (p, w)$ for patient x, the expected utility of x is

$$u(x, C) = p \cdot (v(x) - w).$$

Given a lottery $L = (D, w(p))$, each patient x chooses the contract $C(x) = (p(x), w(p(x)))$ maximizing his or her expected utility. Namely, for each $p \in D$,

$$u(x, C(x)) \ge u(x, (p, w(p))).$$

If there is more than one value of p that maximizes the expected utility of x, we assume that $p(x)$ is the smallest one, so that the cost of serving patient x is minimized. Notice that $p(x)$ depends on x only indirectly, via the function $v(x)$: indeed, $p(x) = p(x')$ whenever $v(x) = v(x')$. Thus we can write $p(x)$ as $p(v(x))$.

As an example, the randomized assignment is a lottery with $D = \{B\}$ and $w(B) = 0$.[16] As another example, any equilibrium assignment is also a lottery, with $D = [0, 1]$ and $w(p)$ always equal to the waiting time of H_1 specified by the equilibrium. Indeed, for every patient x, the contract maximizing that individual's expected utility is to go to the hospital assigned by the equilibrium with probability 1.

Without loss of generality, we assume that D is a subinterval of $[0, 1]$, denoted by $[a, b]$. Indeed, if a patient can choose between $(p_1, w(p_1))$ and $(p_2, w(p_2))$ according to the lottery, then by using a "mixed strategy" he or she can choose to be assigned to H_1 with any probability $p = \alpha p_1 + (1-\alpha)p_2$ with $\alpha \in [0, 1]$, and corresponding expected waiting time $\alpha p_1 w(p_1) + (1-\alpha)p_2 w(p_2)$.

Also, without loss of generality, we assume that the patients' expected waiting time function $p \cdot w(p)$ is convex, and thus differentiable almost everywhere. Indeed, for any contracts $C_1 = (p_1, w(p_1))$, $C_2 = (p_2, w(p_2))$, and $C = (p, w(p))$ with $p = \alpha p_1 + (1-\alpha)p_2$ for some $\alpha \in [0, 1]$, if $p \cdot w(p) > \alpha p_1 w(p_1) + (1-\alpha)p_2 w(p_2)$, then a patient is always better off by mixing between C_1 and C_2 instead of choosing C. Thus we may simply assume that $p \cdot w(p) \le \alpha p_1 w(p_1) + (1-\alpha)p_2 w(p_2)$.[17]

The social welfare and the budget constraint are naturally defined for lotteries as follows.

[16] In general, D can be a proper subset of $[0, 1]$, as the government may not offer the whole interval $[0, 1]$ for the patients to choose from.

[17] Notice that $w(p)$ itself may not be convex.

Definition 8. Given a lottery $L = ([a, b], w(p))$ and the contracts $(p(x), w(p(x)))$ chosen by the patients $x \in [0, 1]$, letting $u(x) \triangleq u(x, (p(x), w(p(x))))$, the *social welfare* of L denoted by SW_L is

$$SW_L = \int_0^1 u(x)\, dx.$$

Lottery L is *feasible* if the budget constraint is satisfied, namely, $\int_0^1 p(x)\, dx = B$.

Notice that we require a feasible lottery to use up all the budget. This is again without any loss of generality, since our theorem below implies that any lottery with cost $B' < B$ is beaten by the randomized assignment with budget B', and thus by the one with budget B.

We assume that the expected waiting time function $pw(p)$ is piecewise twice differentiable in p. Notice that, although assuming twice differentiability of $pw(p)$ over the whole domain is too much, assuming it piecewisely is quite natural. For example, the government may use different $w(p)$'s for different intervals of p, but inside each interval, it uses a smooth $w(p)$. The randomized assignment and equilibrium assignments trivially satisfy this assumption.

The following theorem shows that, when the distribution of the patients' valuations accumulates toward the higher-value side, the randomized assignment is optimal compared with any lottery. Since equilibrium assignments are special cases of lotteries, the randomized assignment is optimal compared with them as well.

Theorem 6. *For any concave valuation function $v(x)$ and any feasible lottery $L = ([a, b], w(p))$, we have $SW_r \geq SW_L$.*

Proof. Since SW_r does not depend on any waiting time, we will rewrite SW_L, so that the waiting times disappear from its representation.

As the choice of $p(x)$ maximizes the utility of x, for any $\Delta > 0$ patient x prefers contract $C(x) = (p(x), w(p(x)))$ to contract $C(x+\Delta) = (p(x+\Delta), w(p(x+\Delta)))$, and patient $x+\Delta$ prefers $C(x+\Delta)$ to C. That is,

$$u(x) = p(x)[v(x) - w(p(x))] \geq p(x+\Delta)[v(x) - w(p(x+\Delta))],$$

and

$$u(x+\Delta) = p(x+\Delta)[v(x+\Delta) - w(p(x+\Delta))] \geq p(x)[v(x+\Delta) - w(p(x))].$$

Accordingly,

$$v(x)\cdot \Delta p(x) \leq \Delta(p(x)\cdot w(p(x))), \qquad \text{and} \qquad v(x+\Delta)\cdot \Delta p(x) \geq \Delta(p(x)\cdot w(p(x))). \tag{13}$$

As $pw(p)$ is piecewise twice differentiable, all the differential equations and statements made in this paragraph hold piecewisely, and we shall not mention the piecewiseness again and again. To begin with, letting $\Delta \to 0$ in Equation (13), we have (with variable x omitted for conciseness)

$$v = \frac{d(pw(p))}{dp}, \tag{14}$$

where the function on the right-hand side is well defined and differentiable in p. As $p(v)$ is the inverse of Equation (14), it is differentiable in v. As $v(x)$ is concave, it is differentiable in x almost everywhere. Thus $p(x) = p(v(x))$ is differentiable in x. Accordingly, we have

$$\begin{aligned} du(x) &= dp\cdot(v-w) + p\cdot(dv - dw) = p\cdot dv + v\cdot dp - (w\cdot dp + p\cdot dw) \\ &= p\cdot dv + v\cdot dp - d(p\cdot w) = p\cdot dv + v\cdot dp - v\cdot dp = p\cdot dv. \end{aligned} \tag{15}$$

(Notice that $p(v)$ and $p(x)$ may not be continuous functions, but we only need them to be "nice" piecewisely.)

Now, putting all the pieces together and integrating both sides of Equation (15) over the whole domain, we have

$$u(x) = \int_0^{v(x)} p(\hat{v})\, d\hat{v}. \tag{16}$$

As $v(x)$ is nondecreasing and concave, we have that $v'(x) \geq 0$ and $v'(x)$ is nonincreasing. If there exists $x < 1$ such that $v'(x) = 0$, then let x_0 be the smallest number with $v'(x_0) = 0$; otherwise (i.e., $v(x)$ is strictly increasing) let $x_0 = 1$. We have that $v(x)$ is strictly increasing on $[0, x_0]$ and constant on $[x_0, 1]$. Let $v_0 = v(x_0)$. Following Equation (16) the social welfare of lottery L is

$$\begin{aligned} SW_L &= \int_0^1 u(x)\,dx = \int_0^1 \int_0^{v(x)} p(\hat{v})\,d\hat{v}\,dx = \int_0^{x_0} \int_0^{v(x)} p(\hat{v})\,d\hat{v}\,dx + \int_{x_0}^1 \int_0^{v_0} p(\hat{v})\,d\hat{v}\,dx \\ &= \int_0^{v_0} \left(p(\hat{v}) \int_{v^{-1}(\hat{v})}^{x_0} dx \right) d\hat{v} + \int_0^{v_0} \left(p(\hat{v}) \int_{x_0}^1 dx \right) d\hat{v} \\ &= \int_0^{v_0} p(\hat{v}) \cdot (x_0 - v^{-1}(\hat{v}))\,d\hat{v} + \int_0^{v_0} p(\hat{v}) \cdot (1 - x_0)\,d\hat{v} \\ &= \int_0^{x_0} p(x)(x_0 - x)v'(x)\,dx + \int_0^{x_0} p(x)(1 - x_0)v'(x)\,dx = \int_0^{x_0} p(x)(1 - x)v'(x)\,dx. \end{aligned} \tag{17}$$

Now that SW_L only depends on the patients' values and their choices of the probabilities, we rewrite SW_r in a similar way so that we can compare the two. Indeed, the social welfare of the randomized assignment can be written as

$$\begin{aligned} SW_r &= \int_0^1 Bv(x)\,dx = \int_0^1 \int_0^{v(x)} B\,dv\,dx = \int_0^{x_0} \int_0^{v(x)} B\,dv\,dx + \int_{x_0}^1 \int_0^{v_0} B\,dv\,dx \\ &= \int_0^{v_0} \int_{v^{-1}(\hat{v})}^{x_0} B\,dx\,d\hat{v} + \int_0^{v_0} \int_{x_0}^1 B\,dx\,d\hat{v} = \int_0^{v_0} B(x_0 - v^{-1}(\hat{v}))\,d\hat{v} + \int_0^{v_0} B(1 - x_0)\,d\hat{v} \\ &= \int_0^{x_0} B(x_0 - x)v'(x)\,dx + \int_0^{x_0} B(1 - x_0)v'(x)\,dx = \int_0^{x_0} B(1 - x)v'(x)\,dx. \end{aligned} \tag{18}$$

Following Equations (17) and (18), to compare SW_r and SW_L, we need to see how the two functions inside the integrals compare to each other, in particular, how $p(x)$ is compared to B. Toward this goal, again notice that $p(x)$ maximizes the expected utility of x. Thus, for any two patients $x_1 < x_2$, we have

$$u(x_1) = p(x_1)(v(x_1) - w(p(x_1))) \geq p(x_2)(v(x_1) - w(p(x_2)))$$

and

$$u(x_2) = p(x_2)(v(x_2) - w(p(x_2))) \geq p(x_1)(v(x_2) - w(p(x_1))).$$

Thus $p(x_2)(v(x_2) - v(x_1)) \geq p(x_1)(v(x_2) - v(x_1))$. If $v(x_2) = v(x_1)$, then $p(x_2) = p(x_1)$ (as we already said, $p(x)$ only depends on $v(x)$), otherwise $p(x_2) \geq p(x_1)$. That is, the function $p(x)$ is *nondecreasing*.

As L is feasible, we have $\int_0^1 p(x)\,dx = B$. Since $v(x)$ is constant on $[x_0, 1]$, so is $p(x)$. Therefore $p(x_0) \geq B$. Accordingly, there exists $x_B \in [0, x_0]$ such that

$$p(x) \leq B \quad \forall\, x < x_B, \qquad \text{and} \qquad p(x) \geq B \quad \forall\, x \geq x_B.$$

Thus we have

$$\begin{aligned} SW_r - SW_L &= \int_0^{x_0} (B - p(x))(1 - x)v'(x)\,dx \\ &= \int_0^{x_B} (B - p(x))(1 - x)v'(x)\,dx + \int_{x_B}^{x_0} (B - p(x))(1 - x)v'(x)\,dx. \end{aligned} \tag{19}$$

Notice that the value of $p(x_B)$ does not affect the value of the integration, thus without loss of generality, we assume $p(x_B) = B$.

For any $x \leq x_B$, we have (a) $B - p(x) \geq 0$, (b) $1 - x \geq 1 - x_B \geq 0$, and (c) $v'(x) \geq v'(x_B) \geq 0$ since $v'(x)$ is nonnegative and nonincreasing, thus

$$(B - p(x))(1 - x)v'(x) \geq (B - p(x))(1 - x_B)v'(x_B) \quad \forall\, x \leq x_B. \tag{20}$$

Similarly, for any $x \geq x_B$, we have (a) $B - p(x) \leq 0$, (b) $0 \leq 1 - x \leq 1 - x_B$, and (c) $0 \leq v'(x) \leq v'(x_B)$, thus

$$(B - p(x))(1 - x)v'(x) \geq (B - p(x))(1 - x_B)v'(x_B) \quad \forall\, x \geq x_B. \tag{21}$$

Combining Equations (19)–(21), we have

$$\begin{aligned}SW_r - SW_L &\geq \int_0^{x_B} (B - p(x))(1 - x_B)v'(x_B)\,dx + \int_{x_B}^{x_0} (B - p(x))(1 - x_B)v'(x_B)\,dx \\ &= (1 - x_B)v'(x_B)\int_0^{x_0} (B - p(x))\,dx.\end{aligned}$$

Since $\int_0^1 p(x)\,dx = \int_0^{x_0} p(x)\,dx + p(x_0)(1 - x_0)$ and $\int_0^1 p(x)\,dx = B = \int_0^{x_0} B\,dx + B(1 - x_0)$, we have

$$\int_0^{x_0} (B - p(x))\,dx = (p(x_0) - B)(1 - x_0).$$

Accordingly,

$$SW_r - SW_L \geq (1 - x_B)v'(x_B)(p(x_0) - B)(1 - x_0) \geq 0,$$

where the second inequality is because $x_B \leq 1$, $v'(x_B) \geq 0$, $p(x_0) \geq B$, and $x_0 \leq 1$.

Therefore, Theorem 6 holds. □

Remark 6. It is worth pointing out that the analysis above holds as long as $(1 - x)v'(x)$ is nonincreasing. Thus the randomized assignment is optimal compared with any lottery even for some convex valuation function such as $v(x) = e^x$. However, since this is still a sufficient condition and it remains unknown whether it is necessary, we choose to state our theorem for concave functions only, which is a well-studied class. It would be very interesting to fully characterize the condition under which the randomized assignment is optimal, but new techniques might be needed for this purpose.

When there are $k > 2$ hospitals, and, in particular, when the patients do not have the same ranking about the hospitals,[18] it becomes much harder to understand the structure of the optimal lottery. One possible way to attack this problem is to consider the "size" of the optimal lottery, namely, how many distributions it should provide for the patients to choose from. For $k = 2$, one can show that any Provision-after-Wait mechanism can be replaced with a menu of only three choices (the cheaper hospital with no waiting, the more expensive hospital with some waiting, or a lottery with some other amount of waiting). For $k > 2$, it might be possible to replace an arbitrary menu of lotteries with one whose size is linear in k or at least is a function of k alone: we leave a detailed study in this direction as a future work.

Acknowledgments. Part of this research was done when the second author was a postdoctoral researcher at the Institute for Advanced Study and the third author was visiting Princeton University. The authors thank Itai Ashlagi for pointing them to important references and several anonymous reviewers and the editor of *Mathematics of Operations Research* for their comments. The first author is supported by the Alfred P. Sloan Fellowship, a National Science Foundation (NSF) CAREER award (CCF-1149888), NSF Award CCF-1215990, a Turing Centenary Fellowship, and a Packard Fellowship in Science and Engineering. The second author is supported in part by the Zurich Financial Services and NSF grant CCF-0832797. The third author is supported by NSF Award CCF-1137084.

References

[1] Aggarwal G, Muthukrishnan S, Pál D, Pál M (2009) General auction mechanism for search advertising. *Proc. 18th Internat. Conf. World Wide Web, WWW '09* (ACM Press, New York), 241–250.

[2] Alatas V, Banerjee A, Hanna R, Olken B, Purnamasari R, Wai-Poi M (2012) Ordeal mechanisms in targeting: Theory and evidence from a field experiment in Indonesia. Technical report, Mimeo (Massachusetts Institute of Technology, Cambridge, MA).

[3] Ashlagi I, Braverman M, Hassidim A (2009) Ascending unit demand auctions with budget limits. Working paper.

[4] Barzel Y (1974) A theory of rationing by waiting. *J. Law Econom.* 17(1): 73–95.

[5] da Graça T, Masson R (2013) Ignorance is bliss? Uncertainty about product valuation may benefit consumers. *Appl. Econom. Lett.* 20(9):897–902.

[6] Dawson D, Gravelle H, Jacobs R, Martin S, Smith PC (2007) The effects of expanding patient choice of provider on waiting times: Evidence from a policy experiment. *Health Econom.* 16(2):113–128.

[7] Demange G, Gale D, Sotomayor M (1986) Multi-item auctions. *J. Political Econom.* 94(4):863–872.

[8] Easley D, Kleinberg J (2010) *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, Chapter 10 (Cambridge University Press, Cambridge, UK).

[9] Felder S (2008) To wait or to pay for medical treatment? Restraining ex-post moral hazard in health insurance. *J. Health Econom.* 27(6):1418–1422.

[10] Gravelle H, Siciliani L (2008a) Is waiting-time prioritisation welfare improving? *Health Econom.* 17(2):167–184.

[18] With two hospitals, as we have discussed, we can assign all patients who prefer the cheaper hospital to it and focus on those who prefer the more expensive one. Thus without loss of generality, we can consider patients who have the same ranking about the hospitals.

[11] Gravelle H, Siciliani L (2008b) Optimal quality, waits and charges in health insurance. *J. Health Econom.* 27(3):663–674.

[12] Gravelle H, Siciliani L (2009) Third degree waiting time discrimination: Optimal allocation of a public sector health care treatment under rationing by waiting. *Health Econom.* 18(8):977–986.

[13] Gravelle H, Sivey P (2010) Imperfect information in a quality-competitive hospital market. *J. Health Econom.* 29(4):524–535.

[14] Hartline J, Roughgarden T (2008) Optimal mechanism design and money burning. *Proc. 40th Annual ACM Sympos. Theory Comput.* (ACM, New York), 75–84.

[15] Heskett J (2003) Shouldice Hospital Limited. Harvard Business School Case 683-068, Boston.

[16] Iversen T (1993) A theory of hospital waiting lists. *J. Health Econom.* 12(1):55–71.

[17] Kahn C, Ault T, Isenstein H, Potetz L, Van Gelder S (2006) Snapshot of hospital quality reporting and pay-for-performance under medicare. *Health Affairs* 25(1):148–162.

[18] Kuhn HW (1955) The Hungarian method for the assignment problem. *Naval Res. Logist. Quart.* 2(1–2):83–97.

[19] Leshno JD (2012) Dynamic matching in overloaded systems. Working paper.

[20] Lindenauer P, Remus D, Roman S, Rothberg M, Benjamin E, Ma A, Bratzler D (2007) Public reporting and pay for performance in hospital quality improvement. *New England J. Medicine* 356(5):486–496.

[21] Lindsay C, Feigenbaum B (1984) Rationing by waiting lists. *Amer. Econom. Rev.* 74(3):404–417.

[22] Ma C, Mak H (2012) Information disclosure and the equivalence of prospective payment and cost reimbursement. Technical report, Department of Economics, Boston University, Boston.

[23] Newhouse JP (2002) *Pricing the Priceless: A Health Care Conundrum*, Chapter 1 (MIT Press, Cambridge, MA).

[24] Rosén P, Anell A, Hjortsberg C (2001) Patient views on choice and participation in primary health care. *Health Policy* 55(2):121–128.

[25] Rosenthal M, Fernandopulle R, Song H, Landon B (2004) Paying for quality: Providers' incentives for quality improvement. *Health Affairs* 23(2):127–141.

[26] Shapley LS, Shubik M (1972) The assignment game I: The core. *Internat. J. Game Theory* 1(2):111–130.

[27] Siciliani L, Hurst J (2005) Tackling excessive waiting times for elective surgery: A comparative analysis of policies in 12 OECD countries. *Health Policy* 72(2):201–215.