Online Appendix for "Efficient Matching under Distributional Constraints: Theory and Applications"

Yuichiro Kamada and Fuhito Kojima

Appendix A. Matching Markets in Practice

A.A. **Residency Matching in Japan.** In Japan, about 8,000 doctors and 1,500 residency programs participate in the matching process each year. This section describes how this process has evolved and how it has affected the debate on the geographical distribution of residents. For further details of Japanese medical education written in English, see Teo (2007) and Kozu (2006). Also, information about the matching program written in Japanese is available on the websites of the government ministry and the matching organizer.[47]

The Japanese residency matching started in 2003 as part of a comprehensive reform of the medical residency program. Prior to the reform, clinical departments in university hospitals, called *ikyoku*, had de facto authority to allocate doctors. The system was criticized because it was seen to have given clinical departments too much power and resulted in opaque, inefficient, and unfair allocations of doctors against their will.[48] Describing the situation, Onishi and Yoshida (2004) write "This clinical-department-centred system was often compared to the feudal hierarchy."

To cope with the above problem a new system, the Japan Residency Matching Program (JRMP), introduced a centralized matching procedure using the (doctor-proposing) deferred acceptance algorithm by Gale and Shapley (1962). Unlike its U.S. counterpart, the National Resident Matching Program (NRMP), the system has no "match variation" (Roth and Peranson, 1999) such as married couples, which would cause many of the good properties of the deferred acceptance algorithm to fail.

Although the matching system was welcomed by many, it has also received a lot of criticisms. This is because some hospitals, especially university hospitals in rural areas, felt that they attracted fewer residents under the new matching mechanism. They argued that

[47]See the websites of the Ministry of Health, Labor and Welfare (http://www.mhlw.go.jp/topics/bukyoku/isei/rinsyo/) and the Japan Residency Matching Program (http://www.jrmp.jp/).

[48]The criticism appears to have some justification. For instance, Niederle and Roth (2003) offer empirical evidence that a system without a centralized matching procedure reduces mobility and efficiency of resident allocation in the context of the U.S. gastroenterologist match.

the new system provided too much opportunity for doctors to work for urban hospitals rather than rural hospitals, resulting in severe doctor shortages in rural areas. While there is no conclusive evidence supporting their claim, an empirical study by Toyabe (2009) finds that the geographical imbalance of doctors has increased in recent years according to several measures (the Gini coefficient, Atkinson index, and Theil index of the per-capita number of doctors across regions). By contrast, he also finds that the imbalance is lower when residents are excluded from the calculation. Based on these findings, he suggests that the matching system introduced in 2003 may have contributed to the widening regional imbalance of doctors.

To put such criticisms into context, we note that the regional imbalance of doctors has been a long-standing and serious problem in Japan. As of 2004, there were over 160,000 people living in the so-called *mui-chiku*, which means "districts with no doctors" (Ministry of Health, Labour and Welfare, 2005b)[49] and many more who were allegedly underserved. One government official told one of the authors (personal communication) that the regional imbalance is one of the most important problems in the government's health care policy, together with financing health care cost. Popular media regularly report stories of doctor shortages, often in a very sensational tone.[50] There is evidence that the sufficient staffing of doctors in hospitals is positively correlated with the quality of medical care such as lower mortality (see Pronovost, Angus, Dorman, Robinson, Dremsizov, and Young (2002) for instance); thus the doctor shortage in rural areas may lead to bad medical care.

In response to the criticisms against the matching mechanism, the Japanese government introduced a new system with regional caps beginning with the matching conducted in 2009. More specifically, a regional cap was imposed on the number of residents in each of the 47 prefectures that partition the country. If the sum of the hospital capacities in a region exceeds its regional cap, then the capacity of each hospital is reduced to equalize the total capacity with the regional cap.[51] Then the deferred acceptance algorithm is

[49]A mui-chiku is defined by various criteria such as the ease of access to hospitals, the population, the regularity of clinic openings, and so forth (Ministry of Health, Labour and Welfare, 2005a).

[50]For instance, the Yomiuri Shimbun newspaper, with circulation of over 10,000,000, recently provoked a controversy by its article about the only doctor in Kamikoani-mura village, where 2,800 people live (Yomiuri Shimbun newspaper, 03/19/2010). Although the doctor, aged 65, took only 18 days off a year, she was persistently criticized by some "unreasonable demanding" patients. When she announced that she wanted to quit (which means that the village will be left with no doctor) because she was "exhausted," 600 signatures were collected in only 10 days, to change her mind.

[51]The capacity of a hospital is reduced proportionately to its original capacity in principle (subject to integrality constraints) although there are a number of fine adjustments and exceptions. Although

implemented under the reduced capacities. We call this mechanism the Japan Residency Matching Program (JRMP) mechanism. The basic intuition behind this policy is that if residents are denied from urban hospitals because of the reduced capacities, then some of them will work for rural hospitals.

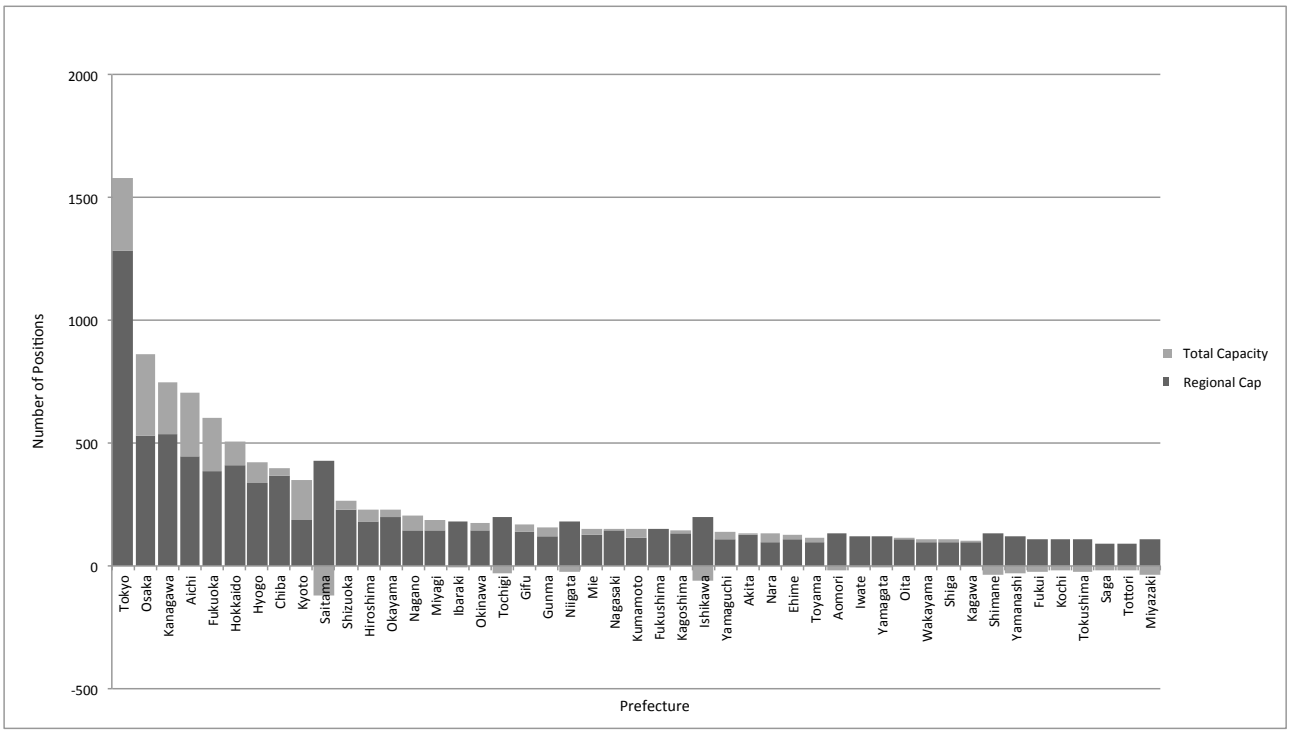

FIGURE 1. For each prefecture, the total capacity is the sum of advertised positions in hospitals located in the prefecture in 2008. The regional caps are based on the government's plan in 2008 (Ministry of Health, Labour and Welfare, 2009a). Negative values of total capacities in some prefectures indicate the excess amount of regional caps beyond the advertised positions.

The magnitude of the regional caps is illustrated in Figure 1. Relatively large reductions are imposed on urban areas. For instance, hospitals in Tokyo and Osaka advertised 1,582 and 860 positions in 2008, respectively, but the government set the regional caps of 1,287 and 533, the largest reductions in the number of positions. The largest reduction in proportion is imposed on Kyoto, which offered 353 positions in 2008 but the number is dropped to 190, a reduction of about 46 percent. Indeed, the projected changes were so

this rule might suggest that hospitals have incentives to misreport their true capacities, the Japanese government regulates how many positions each hospital can offer so that the capacity can be considered exogenous. More specifically, the government decides the physical capacity of a hospital based on verifiable information such as the number of beds in it.

large that the government provided a temporary measure that limits per-year reductions within a certain bound in the first years of operation, though the plan is to reach the planned regional cap eventually. In total, 34 out of 47 prefectures are given regional caps smaller than the numbers of advertised positions in 2008.

The new JRMP mechanism with regional caps was used in 2009 for the first time. The government claims that the change alleviated the regional imbalance of residents: It reports that the proportion of residents matched to hospitals in rural areas has risen to 52.3 percent, an increase of one percentage point from the previous year (Ministry of Health, Labour and Welfare, 2009b).[52] However, there is mounting criticism to the JRMP mechanism as well. For instance, a number of governors of rural prefectures (see Tottori Prefecture (2009) for instance) and a student group (Association of Medical Students, 2009) have demanded that the government modify or abolish the JRMP mechanism with regional caps.[53] Among other things, a commonly expressed concern is that the current system with regional caps causes efficiency losses, for instance by preventing residents from learning their desired skills for practicing medical treatments.

In the main text we formalized the JRMP mechanism (Section II), explored its properties (Example 1, Remark 1, and Proposition 4), and compared it to the flexible deferred acceptance mechanism (Theorem 3). Our analysis suggests that the current JRMP mechanism needs to be changed to the flexible deferred acceptance mechanism.

A.B. **Chinese Graduate Admission.** This section describes the Chinese graduate admission in detail, and formally shows that the mechanism may result in an unstable and inefficient matching.[54]

A.B.1. *Institutional Background.* Chinese society is changing rapidly, and it is widely believed that there is need for more workers with professional master's degrees.[55] However, professional master's degrees have traditionally been regarded as inferior to academic

[52]Ministry of Health, Labour and Welfare (2009b) defines "rural areas" as all prefectures except for 6 prefectures, Tokyo, Kyoto, Osaka, Kanagawa, Aichi, and Fukuoka, which have large cities.

[53]Interestingly, even regional governments in rural areas such as Tokushima and Tottori opposed to the JRMP mechanism. They were worried that since the system reduces capacities of *individual hospitals* in the region and some of them could hire more residents, it can reduce the number of residents allocated in the regions even further. This feature - inflexibility of the way capacities are reduced - is one of the problems of the current JRMP mechanism that we try to remedy by our alternative mechanism.

[54]We greatly benefited from discussing Chinese graduate school admission with Jin Chen.

[55]Ministry of Education of China (2010) states that "[the education authority and graduate schools] should put emphasis on the promotion of education for advanced professionals, especially full-time professional master's degree."

master's degrees by many.[56] And there are not as many students in professional master's programs as the government aims.[57]

To address this issue, Chinese government started a new regulation to increase enrollment in professional master's programs in 2010 (People's Republic of China, 2010). More specifically, the government began to impose constraints on the total number of academic master students, while increasing the number of professional master students. To achieve this goal, the government decided to reduce the available seats of each academic master's program by about 25 percent by 2015, while increasing capacities for professional programs.[58]

Although Chinese graduate school admission is different from Japanese residency match in many ways, there is a clear isomorphism between the structures of the problems that these markets are faced with. Just as there is demand for increasing resident allocation in rural Japan, there is demand for increasing professional master students in Chinese graduate education. Moreover, in both Japanese and Chinese cases, the feasibility requirements are placed on the total numbers of allocations for a subset of institutions (hospitals in each prefecture in the Japanese case, and the academic master's programs in the Chinese case). Lastly, when implementing the requirement, both governments place rigid restrictions on the allowed seats of each institution (each hospital in Japan, and each graduate school in China).

Remark 4. As mentioned in Section V.A.1 of the main text, the matching mechanism for Chinese graduate school admission has additional rounds. First, there is a recommendation-based admission for high achievers before the main round. Second, there is a round called an "adjustment process" for those who have not been matched by the

[56]Chinese government has been emphasizing that the only differences between professional and academic master's programs are in the types and goals of education, and not in standards, but the reputation of a professional master's degree is still not as good as an academic one (Zhai, 2011).

[57]China's master-level education system traditionally emphasized academic training, rather than professional training. However, most graduates from master's programs pursue a professional career instead of an academic one: in 2009, for instance, enrollment for master's programs was around 415,000 while that for PhD programs was around 60,000 (in China, the master's degree is similar to M.Phil. in countries like the U.K. in that a student seeking PhD first attend a master's program). Between 2008 and 2011, the proportion of professional masters' enrollment has increased from just 7 percent to 30 percent of the total enrollment. Ministry of Education aims at 50 percent by 2015 (Lin, 2011).

[58]To achieve this goal gradually, the government plans to reduce the number of seats by about 5 percent every year until 2015 (to our knowledge, the government has not disclosed whether it will continue imposing the reduction beyond 2015).

end of the main round. We do not formally analyze these rounds because they are not directly related to the issue of distributional constraints, and processes similar to them and the associated problems have been analyzed by other works.[59] Nevertheless, we describe these rounds for completeness.

In the recommendation-based admission, students are recommended to graduate schools directly even before the graduate entrance examination that everyone else should take for admission. This round works as a shortcut for excellent students to enter graduate schools. Most recommended students are admitted to and attend the same university's graduate school where he or she attended college, and few of the recommended students transfer to another graduate school. This round is decentralized, and how to evaluate and admit students is largely up to each school, and thus the admission policy varies from school to school.

Students who were unmatched in the recommendation-based admission and the main round, as well as graduate schools which were not full in these rounds, enter the adjustment process. This round proceeds in real time, and each student maintains an application list which consists of at most two schools at any moment during this process. More specifically, at the beginning of this round, each student enters at most two schools into an online system.[60] Each school sees students who listed it, decides whether to give each applicant the chance for interview or not, and sends out interview invitations to them. Upon receiving an interview invitation, a student chooses whether to accept it or not. If a student fails to receive an interview invitation in 48 hours or is rejected after the interview from a school, the student can remove the school and add another school into her online application list. Once changed, a student must keep the new school in the list for at least 48 hours unless it interviews her. On the interview day, a graduate school admits or rejects students. If the capacity becomes full, the graduate school completes the admission. Each student can confirm admission from at most one school. Once she confirms, she exits the matching process.

[59]Abdulkadiroğlu, Pathak, and Roth (2005); Abdulkadiroğlu, Pathak, and Roth (2009) study New York City's high school match. In NYC, top 2 percent students are automatically admitted to certain schools if they prefer, similarly to top performers who can be admitted to some schools in China's recommendation-based admission. Roth and Xing (1994, 1997) study labor markets that proceed in real time, and highlight the time constraint and associated strategic behavior and inefficiency of the resulting matching.

[60]The application is maintained at http://yz.chsi.com.cn/.

The adjustment process happens in real time: In 2012, for instance, the process was in session from April 1st to May 5th. Given this time constraint, it is widely believed that students and schools contact each other ahead of the official adjustment process.[61] □

A.B.2. *Formal Analysis.* Let us use the same notation as in the main text, although now we call h a program instead of a hospital, and d a student instead of a doctor. Further assume that the set of all programs is partitioned into the set of all academic programs r and the set of all professional programs r'. Throughout, assume $q_{r'} > \sum_{h \in H_{r'}} q_h$ so that the cap for professional programs is not binding.

We define the main round of the Chinese graduate admission formally.[62] As described in the main text, given the cap q_r, the main round of Chinese graduate admission runs as follows (we describe the mechanism for a general value of q_r, although in China q_r is an integer close to 75 percent of the sum of academic program capacities). Set $\bar{q}_h \leq q_h$ for each h in such a way that $\sum_{h \in H_r} \bar{q}_h \leq q_r$ (in Chinese graduate admission, $\bar{q}_h$ is an integer that is at most 75 percent of q_h for each $h \in H_r$). Each student applies to at most one program. Given the set of applicants, each program h accepts its most preferred students up to its target capacity $\bar{q}_h$ and rejects everyone else. All matchings are final.

We model the behavior in this mechanism by considering the following two-stage extensive-form game. In the first stage, students simultaneously apply to programs, one for each student. Then in the second stage, each program admits students from those who applied to it up to its target capacity. In this game, the following result holds:

Result 1. *Suppose that $q_r > \sum_{h \in H_r} q_h$ and $\bar{q}_h = q_h$ for all h. Then the set of the pure-straetegy subgame-perfect equilibrium outcomes in the game induced by the Chinese graduate admission coincides with the set of stable matchings.*

Proof. When $q_r > \sum_{h \in H_r} q_h$ and $\bar{q}_h = q_h$ for all h, the stability concept of this paper is equivalent to the standard stability concept (as in Roth and Sotomayor (1990) for example). By Sotomayor (2004) and Echenique and Oviedo (2006), the set of subgame perfect equilibrium outcomes of this game is equivalent to the set of stable matchings in the standard sense. These two observations complete the proof. □

[61] See "Experts: 2012 graduate school entrance examination, advice and strategies for adjustment" from *Kuakao Education*, a counseling agency in China for students applying for graduate schools, available at http://yz.chsi.com.cn/kyzx/fstj/201203/20120322/293594478.html.

[62] The description is based on the website of the "National Graduate Admissions Information Network" (http://yz.chsi.com.cn/), which provides information on graduate admission and host online applications for graduate schools.

Thus if there is no binding cap on academic programs, then the equilibrium outcomes are stable and hence efficient. When the cap on academic programs is binding as in the Chinese admission mechanism, however, neither of these good properties hold even for equilibrium outcomes. The following example, which is an adaptation of Example 1, is such a case.

Example 10 (Equilibrium Outcomes under the Chinese Mechanism Can Be Unstable and Inefficient)**.** The "regional cap" for academic programs r is $q_r = 10$. There are two academic programs h_1 and h_2 and no professional program. Each program h has a capacity of $q_h = 10$. Let the target capacities be $\bar{q}_{h_1} = \bar{q}_{h_2} = 5$. There are 10 students, $d_1, \dots, d_{10}$. Preference profile $\succ$ is as follows:

$$\succ_{h_i}\colon d_1, d_2, \dots, d_{10} \text{ for } i = 1, 2,$$
$$\succ_{d_j}\colon h_1 \text{ if } j \leq 3 \quad \text{and} \quad \succ_{d_j}\colon h_2 \text{ if } j \geq 4.$$

It is easy to see that the only pure-strategy subgame perfect equilibrium outcome is

$$\mu = \begin{pmatrix} h_1 & h_2 & \emptyset \\ d_1, d_2, d_3 & d_4, d_5, d_6, d_7, d_8 & d_9, d_{10} \end{pmatrix}.$$

Consider a matching μ' defined by,

$$\mu' = \begin{pmatrix} h_1 & h_2 \\ d_1, d_2, d_3 & d_4, d_5, d_6, d_7, d_8, d_9, d_{10} \end{pmatrix}.$$

Since the cap for academic programs is still respected, μ' is feasible. Moreover, every student is weakly better off with students d_9 and d_{10} being strictly better off than at μ. Hence we conclude that the Chinese mechanism can result in an inefficient matching. We also note that μ is not stable: For example, program h_2 and student d_9 constitute a blocking pair while the cap for r is not binding. □

A.C. **College Admission in Ukraine.** A problem similar to Japanese residency match and Chinese graduate school admission is found in college admission in Ukraine as well. In Ukraine, some of the seats are financed by the state, while other "open-enrollment" seats require that students pay tuition (Kiselgof, 2012).[63] There is a cap on the number of state-financed seats, apparently as there is a limit on the budget that can be used to finance college study. The government implements the cap on the number of state-financed seats

[63]A similar cap on the number of state-financed college seats exists in Hungarian college admission as well. The situation is somewhat different here, however, as ranking by colleges are based on a common exam and hence is common for different university programs on the same subject. Biró, Fleiner, Irving, and Manlove (2010) propose an elegant matching mechanism in such an environment.

by imposing a cap on each program as in Japanese residency match and the Chinese graduate admission. Although the specific mechanism the Ukrainian college admission system uses is different from JRMP and Chinese graduate school admissions (see Kiselgof (2012) for detail), instability and inefficiency because of the constraints similarly result.

A.D. **Medical Matching in the United Kingdom.**

A.D.1. *Institutional Backgrounds.* In recent years, how to organize medical training has been a contentious topic in the U.K., and the system has undergone a number of drastic changes. This section describes the current system, whose basic structure was set up in 2005.[64]

In order to practice medicine in the U.K., graduates from medical schools must undertake two years of training. The arrangement is called the Foundation Programme, and places about 7,000 medical school graduates to training programs every year.[65] In the first round of the matching scheme of the Foundation Programme, applicants are matched to one of 25 "foundation schools" by a national matching process. A foundation school is a consortium made of medical schools and other organizations, and each foundation school largely corresponds to a region of the country. Upon being matched to a foundation school, students are matched to individual training programs within that foundation school in the second round of the matching process. In these processes, applicants are assigned a numerical score, which may result in ties. Until 2011, the Boston mechanism (also known as the "first-choice-first" mechanism in the U.K.) was in use, based on the numerical score and a random tie-breaking.[66] Beginning in 2012, a serial dictatorship algorithm based on the score with random tie-breaking is used for allocation to foundation schools.[67] It is up to individual foundation schools as to how they match their assigned applicants to programs in their region. In Scotland, for example, a stable mechanism was

[64]We are grateful to Peter Biró, Rob Irving, and David Manlove for answering our questions about medical match in the U.K.

[65]Some institutional details and the statistics reported here can be found in the Foundation Programme's webiste, especially in its annual reports: see for example its 2011 annual report at http://www.foundationprogramme.nhs.uk/download.asp?file=Foundation_Programme_Annual_Report_Nov11_FINAL.pdf.

[66]See Abdulkadiroğlu and Sönmez (2003) who study the Boston mechanism in the school choice context.

[67]The algorithm is described at http://www.foundationprogramme.nhs.uk/pages/medical-students/faqs#answer39

in use to allocate students to programs within the region until serial dictatorship based on applicant scores and random tie-breaking replaced it in 2010.[68]

For our purposes, an especially interesting point is that the mechanism used in U.K. medical matching has two rounds, in which students are assigned to a region first, and then to a program within their assigned region. Although the specific mechanisms used in the second round vary from region to region, the United Kingdom as a whole uses a two-round mechanism.

A.D.2. *Formal Analysis.* As indicated above, mechanisms used in U.K. medical match (and Scottish teacher matching as mentioned in Appendix A.E) have many variations, but their basic structure is common in the sense that applicants are matched by a two-round procedure. Formally, we consider a mechanism in which applicants are matched to a region (up to its regional cap) in the first round, and then they are matched to a hospital within the assigned region in the second round. For concreteness we focus on the mechanism in which serial dictatorship is used in both rounds, but the main conclusions can be obtained for other mechanisms as well (we describe the details later in this section).

The first example shows that the outcome of this two-round mechanism may be unstable.

Example 11. There are two regions r_1 and r_2 with regional caps $q_{r_1} = q_{r_2} = 2$. There are two hospitals h_1 and h_3 in r_1 while there is one hospital h_2 in r_2. Hospital capacities are $q_{h_1} = 1$, $q_{h_2} = 2$, and $q_{h_3} = 1$. Suppose that there are 3 doctors, d_1, d_2, and d_3. Preference profile $\succ$ is as follows:

$$\succ_{h_i}\colon d_1, d_2, d_3 \text{ for all } i,$$
$$\succ_{d_j}\colon h_1, h_2, h_3 \text{ for all } j.$$

And let us assume that, in both rounds, the serial dictatorship is used with respect to the ordering d_1, d_2, and d_3. That is, d_1 is matched to her most preferred region (or hospital), d_2 is matched to the most preferred region (or hospital) that are still available, and so on. Note that we assume that hospital preferences are common and coincide with the applicant ordering in the serial dictatorship. This assumption is meant to make stability as easy to obtain as possible, because if serial order and hospital preferences are different, it is almost trivial to obtain unstable matchings (indeed, under the original serial dictatorship,

[68]There are applicants who participate as couples, and the algorithms handle these couples in certain manners. See Irving and Manlove (2009) for details.

the resulting matching is stable under this assumption, but not otherwise). Assume that each docotor prefers r_1 most and r_2 second.[69] Let target capacities be arbitrary.

In the first round of this mechanism, d_1 and d_2 are matched to r_1, while d_3 is matched to r_2. In the second round, d_1 is matched to her first choice h_1, d_2 is matched to h_3, which is the only remaining hospital in region r_1, and d_3 is matched to hospital h_2, resulting in

$$\mu = \begin{pmatrix} h_1 & h_2 & h_3 \\ d_1 & d_3 & d_2 \end{pmatrix}.$$

This matching μ is unstable, because d_2 and h_2 form a legitimate blocking pair: d_2 prefers h_2 to its match h_3 and h_2 prefers d_2 to its match d_3. □

Although we phrased the above example in the context of a mechanism both of whose rounds employ the serial dictatorship, the same point can be made for other two-round mechanisms. Consider, for instance, the case in which the first-round procedure is a Boston mechanism (as was the case in the U.K. until 2011) based on the above ordering, while the second round is a serial dictatorship.[70] In the above market, under this procedure d_1 and d_2 are still matched to region r_1 in the first round, and then d_2 is matched to h_3 in the second round, leading to the same matching μ of Example 11, thus to instability. This observation shows that the problem of instability is not restricted to the detail of the current mechanism using serial dictatorship in both rounds, but rather a general feature of two-round systems that have been the basic framework of the U.K. medical match.

The following example shows that the matching resulting from the U.K. medical match can be inefficient.

Example 12. There are two regions r_1 and r_2 with regional caps $q_{r_1} = 2$ and $q_{r_2} = 1$. There are two hospitals h_1 and h_3 in r_1 while there is one hospital h_2 in r_2. Each hospital h has a capacity of $q_h = 1$. Suppose that there are 3 doctors, d_1, d_2, d_3. Preference profile

[69] Such reported preferences may arise if, for instance, she believes that there is nonzero probability to be matched with h_1 and her cardinal utility from h_1 is sufficiently high.

[70] Recall that the first round algorithm was changed from the Boston mechanism to the serial dictatorship only beginning in 2012 in the U.K. medical match, while serial dictatorship was already in use in Scotland in 2009.

$\succ$ is as follows:

$$\succ_{h_i}\colon d_1, d_2, d_3 \text{ for all } i,$$
$$\succ_{d_1}\colon h_1, h_2, h_3,$$
$$\succ_{d_2}\colon h_1, h_2,$$
$$\succ_{d_3}\colon h_1, h_3, h_2.$$

As before, let us assume that the serial dictatorship with ordering d_1, d_2, and d_3 is used in both rounds. Assume further that doctors' preferences over the regions are induced in the manner specified in Example 11.

At the first round of this mechanism, d_1 and d_2 are matched to r_1, while d_3 is matched to r_2. In the second round, d_1 is matched to her first choice h_1, d_2 is unmatched, and d_3 is matched to hospital h_2, resulting in matching

$$\mu = \begin{pmatrix} h_1 & h_2 & h_3 & \emptyset \\ d_1 & d_3 & \emptyset & d_2 \end{pmatrix}.$$

Consider a matching μ' defined by,

$$\mu' = \begin{pmatrix} h_1 & h_2 & h_3 \\ d_1 & d_2 & d_3 \end{pmatrix}.$$

The latter matching satisfies all regional caps and Pareto dominates the former matching μ: d_1 and h_1 are indifferent between μ and μ', while every other agent is made strictly better off at μ' than at μ. Therefore the matching μ is inefficient. □

The last drawback of the two-round mechanism we point out involves incentives. Serial dictatorship is strategy-proof, and this property is often regarded as one of the main advantages of this mechanism. In a two-round mechanism, however, there exists no dominant strategy even if both rounds employ serial dictatorship.

Example 13. Consider the market defined in Example 11. Note that it is a weakly dominant strategy to report true preferences in any subgame of the second round, i.e., once doctors are matched to regions, so assume that all doctors report true preferences in the second round. Suppose that doctors report preferences over regions as in Example 11. Then doctor d_2 is assigned to her third choice hospital h_3. However, if d_2 reports region r_2 to be her most preferred region while no other doctor changes his reported preference, then she is matched to h_2, which is the optimal matching possible for any of her reported preferences. In other words, reporting r_2 as the most preferred region is a best response while reporting r_1 is not. Next, consider a report of d_1 that reports r_2 to be his most

preferred region. Then d_2 is matched to h_1 if she reports r_1 to be her most preferred region while she is matched to her less preferred hospital h_2 if she reports r_2 to be her most preferred region. In other words, reporting r_1 as the most preferred region is a best response while reporting r_2 is not. Therefore there is no dominant strategy.[71] □

A.E. **Probationary Teacher Matching in Scotland.** Another problem of interest is the matching of new teachers (called probationary teachers) to schools. Teachers in Scotland need to get a training as probationers for one year. The General Teaching Council for Scotland (GTCS) runs a procedure called the Teacher Induction Scheme, which allocates probationary teachers to training posts in Scottish schools.[72] Scotland has 32 local authorities, and probationary teachers and these local authorities are matched in the first round of the mechanism. Information about the algorithm used is unavailable to our knowledge, but some documents suggest that a slight variant of the random serial dictatorship is used.[73] Then each local authority decides which probationers matched to it are sent to which schools under its control, and that round occurs subsequently to the first round. The mechanism that local authorities use in this round is up to each local authority, and appears to vary widely from one local authority to another.

This scheme has a lot in common with the previous example. As in the U.K. medical match, the matching clearinghouse first assigns teachers to a local authority, who then assigns them to schools under its control.

Appendix B. Simulation Methods and Results for the case of the Japanese Residency Matching Program

In this appendix we provide results of simulations using the data on Japanese medical residency match. As we have discussed in the main text of the paper (and in the appendix), there are a number of instances around the world where distributional constraints are

[71] It is trivial, and hence omitted, to show that reporting no region to be acceptable is weakly dominated, so the above argument is enough to establish the claim.

[72] See http://www.gtcs.org.uk/home/students/teacher-induction-scheme-faq.aspx.

[73] "Teacher Induction Scheme 2008/2009," http://www.scotland.gov.uk/Resource/Doc/200891/0053701.pdf states that "A computer system will match and allocate students to local authorities using each local authority's vacancy list and student's preference list. You will be chosen at random and matched against your five preferences, beginning with your first preference. Where an appropriate vacancy is unavailable, you will be matched against your second preference, and so on until an appropriate match is found." As indicated above, a probationary teacher is asked to only rank 5 local authorities, unlike the exact random serial dictatorship. Another complication is that a student can alternatively tick a preference waiver box indicating that they are happy to work anywhere in Scotland. Those who choose the preference waiver option are paid additional compensation.

imposed. We chose the Japanese case as our principal target for simulation for various reasons: First, all the data we used in our simulation are available online for free, so one can replicate our simulation results easily. Second, the Japanese medical match is highly centralized, so the conclusions from the simulation results are more meaningful than in some other applications where a (sometimes nontrivial) part of the matching procedure is not explicitly specified. Third, compared to the current practice (the JRMP mechanism), our proposal (the flexible deferred acceptance (FDA) mechanism) has an advantage in terms of efficiency and stability, while its effect on the regional balance of doctors is ambiguous from the theoretical perspective. Thus simulations are useful. Finally, we have been talking to the Japanese government officials about the FDA mechanism, and by quantifying the trade-off that we have just mentioned, we have a better chance of persuading them to use the FDA mechanism. Since the practicality of the theory is one of our main goals, we view simulations as a useful analysis to implement.

As we reviewed in the paper, the JRMP mechanism was introduced in 2009, and after that, hospitals have been asked to gradually decrease their respective capacities, to eventually match the total capacities to the planned regional cap. Since the government publicizes only the reduced capacities, we use the hospital capacities in the data just before 2009. More specifically, we use the data from 2007, because the government specified the regional cap based on that year's data. For consistency, we also use other parts of information from the data of the same year. All the data used here can be obtained at the webpage of Japanese Medical Residency Program (http://www.jrmp.jp). The data are in Japanese.

B.A. **Simulation Method.** We obtain the data of regional caps of all 47 prefectures, the capacity and the region (prefecture) of each hospital.

Using these data, we set the "target capacity" for each hospital following the description in footnote 13 of the paper. That is, if the sum of the advertised positions in the hospital's region is no more than the regional cap, then the hospital's target capacity is equal to the number of its advertised positions; Otherwise, the target is given by the advertised number times the fraction of the regional cap over the total number of the advertised positions in the region.[74]

The market size. In the simulation, the numbers of doctors and hospitals are 8,291 and 1,357, respectively, which are the number of doctors and hospitals that actually submitted preference lists in the Japanese residency match in 2007.

[74]If the resulting number is not an integer, then we round the numbers to one of the adjacent integers in such a way that the sum of the target capacities in the region is equal to the regional cap.

Preference lists. We do not have the actual data on submitted or true preferences of the doctors and hospitals, so given the above information we generated preferences and ran the simulation. Fortunately, various public data enabled us to set parameters that mimic the Japanese case, as we explain below.

(1) **Doctors.** We obtain the data on the distribution of the length of preference lists of doctors (i.e., the number of hospitals listed in the submitted preference list of each doctor) up to length 8. In the data, the number of doctors who listed k hospitals is not available for $k \geq 9$, while the total number of doctors who listed 9 or more hospitals is available. We also obtain from the data the average number of hospitals listed, which is 3.48.

For the doctors who list 9 or more hospitals in their preference list, we used the truncated exponential distribution such that (i) the number of doctors who list k hospitals is 0.6 times the number of doctors who list $k-1$ doctors, for $k = 10, 11,$ (modulo integer constraints) (ii) the number of doctors at length 9 is adjusted so that the average number of listed hospitals is 3.48, the average from the data, and (iii) the maximum length is 15. With this specification the number of doctors at length 9 is smaller than the number for length 8.

The data describe, for each hospital, the number of doctors who listed that hospital in their preference lists. Using these data, for each hospital h, we define p_h to be the number of doctors who listed it in their preference list divided by the sum of all those numbers across all hospitals (so that it becomes probability, i.e., the numbers sum up to one). Then each doctor with preference list length k independently draws hospitals based on this distribution $(p_h)_{h \in H}$, repeatedly k times without replacement, listing her first pick as the first choice, her second pick as the second choice, and so on.[75]

(2) **Hospitals.** Each hospital ranks doctors uniform randomly, viewing every doctor acceptable.[76]

Remark 5. There could be alternatives for this method. For example we could have each hospital always rank a doctor with a shorter preference-list length higher than the one with a longer length and those who have the same length are ranked

[75]This manner of preference generation is used in a number of matching papers, such as Kojima and Pathak (2009).

[76]In our simulation code, each hospital actually orders only doctors who find the hospital acceptable. This is without any consequence because none of the algorithms we consider in this paper is affected by whether a doctor who finds a hospital unacceptable is acceptable to the hospital.

uniformly randomly: this might as well be closer to the data, because those doctors who rank only a small number of hospitals may be doing so because they are confident that hospitals rank them high. However, without better data on the doctor preference, we did not have better foundation for conducting such biased data generations, and hence stuck to the uniformly-random data generation. We hope that the Japanese government disclose anonymous data of preference lists.

B.B. **Simulation Results.** Medical matching has two sides, namely doctors and hospitals. In the context of the Japanese medical match, another important issue is the distributional balance of doctors in different regions. Therefore, in the following we discuss the simulation results pertinent to welfare of doctors and hospitals, and then discuss the distributional consequences of the mechanisms across regions.

B.B.1. *Doctors and Hospitals.*

(1) **The number of matched doctors.** Figure 2 shows that the cost of using the JRMP mechanism is quite significant: almost 600 doctors who are matched in the unconstrained deferred acceptance (DA) mechanism become unmatched in JRMP (1396 versus 805). However, much of the negative effects can be alleviated if we switch to FDA: the number of additional doctors who become unmatched compared to the DA is about 205 (1010 versus 805), which is only about one third of 600, the corresponding number for the JRMP mechanism. Importantly, FDA achieves this improvement while satisfying all the regional caps just like the JRMP does, while DA does not respect regional caps.

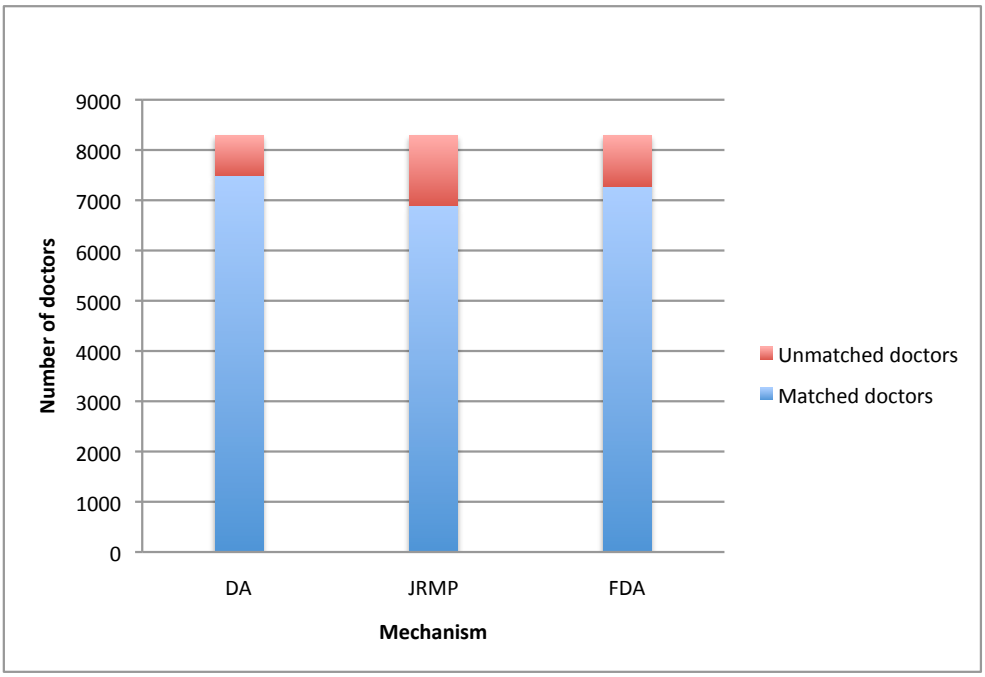

FIGURE 2. The numbers of matched doctors under different mechanisms

(2) **The number of doctors who are made strictly better off from JRMP to FDA, from FDA to DA, and from JRMP to DA.**

From/To	DA	FDA	JRMP
DA	0	0	0
FDA	606 (7.3 %)	0	0
JRMP	1547 (18.7 %)	996 (12.0 %)	0

TABLE 1. The number of doctors who are made strictly better off from JRMP to FDA, from FDA to DA, and from JRMP to DA.

Table 1 demonstrates the numbers of doctors who become strictly better off by changing the mechanism from the one in the row to the one in the column. For example, 996 doctors become strictly better off by changing the mechanism from the JRMP mechanism to the FDA mechanism. As Theorem 3(1) predicts, there are no doctors who become strictly better off from FDA to JRMP, or from DA to the other two mechanisms, and this prediction is confirmed by the zeros in Table 1.

The theorem predicts that doctors are *weakly* better off by the change from JRMP to the other two mechanisms and from FDA to DA, but it does not pin

down how many doctors become *strictly* better off, and in general it is hard to obtain an analytical result on strict improvement without making additional assumptions on the preference distributions. This is one of the main motivations for our simulations. Since DA is unconstrained with respect to the number of doctors that can be matched to each region, large magnitudes of improvement from the other two mechanisms to DA is expected. Even so, the simulation result shows that the improvement from FDA to DA is moderate (7.3%). The result also shows that the effect of the change from JRMP to FDA–both of which are constrained by the regional cap– is large (12.0%). In view of the fact that the DA gives a (loose) upper bound of what FDA can possibly achieve, the simulation result demonstrates FDA's surprisingly large improvements in doctor welfare upon the current JRMP mechanism.

(3) **Cumulative number of doctors matched to their k-th or better choices**

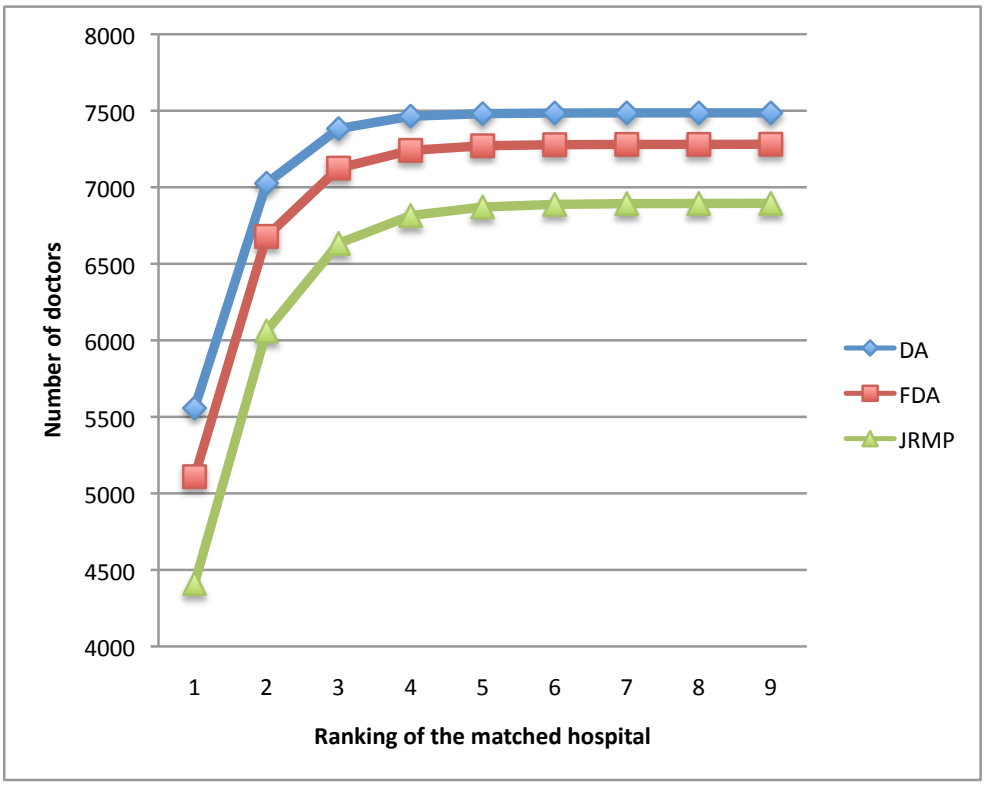

FIGURE 3. Cumulative number of doctors matched to their k-th or better choices.

In the above graph, the horizontal axis describes the ranking, and the vertical axis describes the number of doctors. For each of the mechanisms, we plot the

cumulative number of doctors who are matched to their k-th or better choices, for each value k in the horizontal axis.[77]

The graph confirms our prediction in Theorem 3(1). That is, the doctors are better off under DA than under FDA and under FDA than under JRMP. As we discussed in (2), the theorem does not predict magnitudes of the improvement, and this motivates simulations. Although regional caps certainly result in worse outcomes for doctors, more than a half of the loss caused by JRMP compared to the unconstrained DA can be avoided once FDA is used, even though regional caps are satisfied in FDA just as in JRMP. The effect is large: for example, about 500 more doctors are matched to their first choices under FDA compared to JRMP.

(4) **The number of hospitals that are matched to more doctors in DA, FDA, and JRMP.**

From\To	DA	FDA	JRMP
DA	0	138 (10.2%)	222 (16.4%)
FDA	104 (7.7 %)	0	158 (11.6 %)
JRMP	366 (27.0 %)	376 (27.7 %)	0

TABLE 2. The number of hospitals that are matched to more doctors in DA, FDA, and JRMP.

The above table describes the number of the hospitals that gained more doctors in one mechanism than another. For example, 376 hospitals (27.7 % of the total of 1,357 hospital programs) are matched with more doctors in FDA than in JRMP. Unlike the corresponding table for the doctors, (2), some hospitals receives additional doctors while others lose doctors in any transition between the 3 mechanisms. But overall, the number of hospitals that receive more doctors is larger than the number of hospitals that lose doctors in transitions from JRMP to FDA or DA, which is not surprising given that more doctors are matched in FDA and DA than in JRMP (although this does not necessarily imply that more hospitals are matched under FDA or DA than under JRMP).

B.B.2. *Regions.* It is not surprising that improvement happens for some prefectures from JRMP or FDA to DA. Also, more than 3/4 of the prefectures are assigned more doctors

[77]In this figure, we plot only the doctors who are matched with some hospital. This is because it is the information that JRMP provides in their reports (and that appears to be reasonable statistics which most people care about).

under FDA than under DA. This indicates that, due to the introduction of regional caps, many prefectures became better off in terms of the number of doctors. The more relevant question is the comparison of the improvements from JRMP to FDA and from FDA to JRMP. The exact comparison of the numbers may not make too much sense, but the numbers indicate that the introduction of FDA does not create a situation where "most regions get worse off." This is one finding that the theory did not tell us.

From\To	DA	FDA	JRMP
DA	0	37 (78.7%)	30 (63.8%)
FDA	6 (12.8 %)	0	25 (53.2 %)
JRMP	16 (34.0 %)	20 (42.6 %)	0

TABLE 3. The number of regions that are assigned strictly more doctors from JRMP to FDA, from FDA to DA, and from JRMP to DA.

An issue of interest that is suppressed in the above table is the magnitude of improvements and decline in different regions.

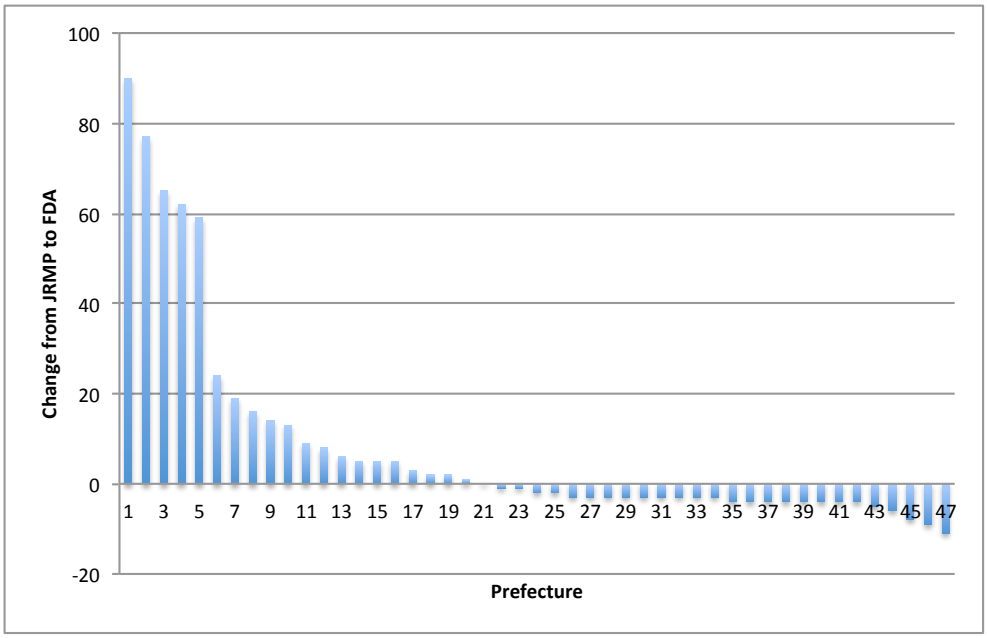

FIGURE 4. The magnitude of improvements and decline in different regions.

To better see such magnitude, in the above figure we plot the change in the number of doctors in each region, in a descending order. The graph shows that the magnitude of improvement from JRMP to FDA is large while that of decline is small. To take some

numbers, the maximum improvement is 90 while the maximum decline is 11. The area above the positive region is larger than the one in the negative region, which is consistent with our overall finding that FDA assigns about 400 more doctors in hospitals than JRMP does.

This graph is, however, silent about the distributional consequences across regions. This issue appears to be one of the main concerns in Japan. To study this issue, the figure below plots which regions become better off and which become worse off in the transition from JRMP to FDA.

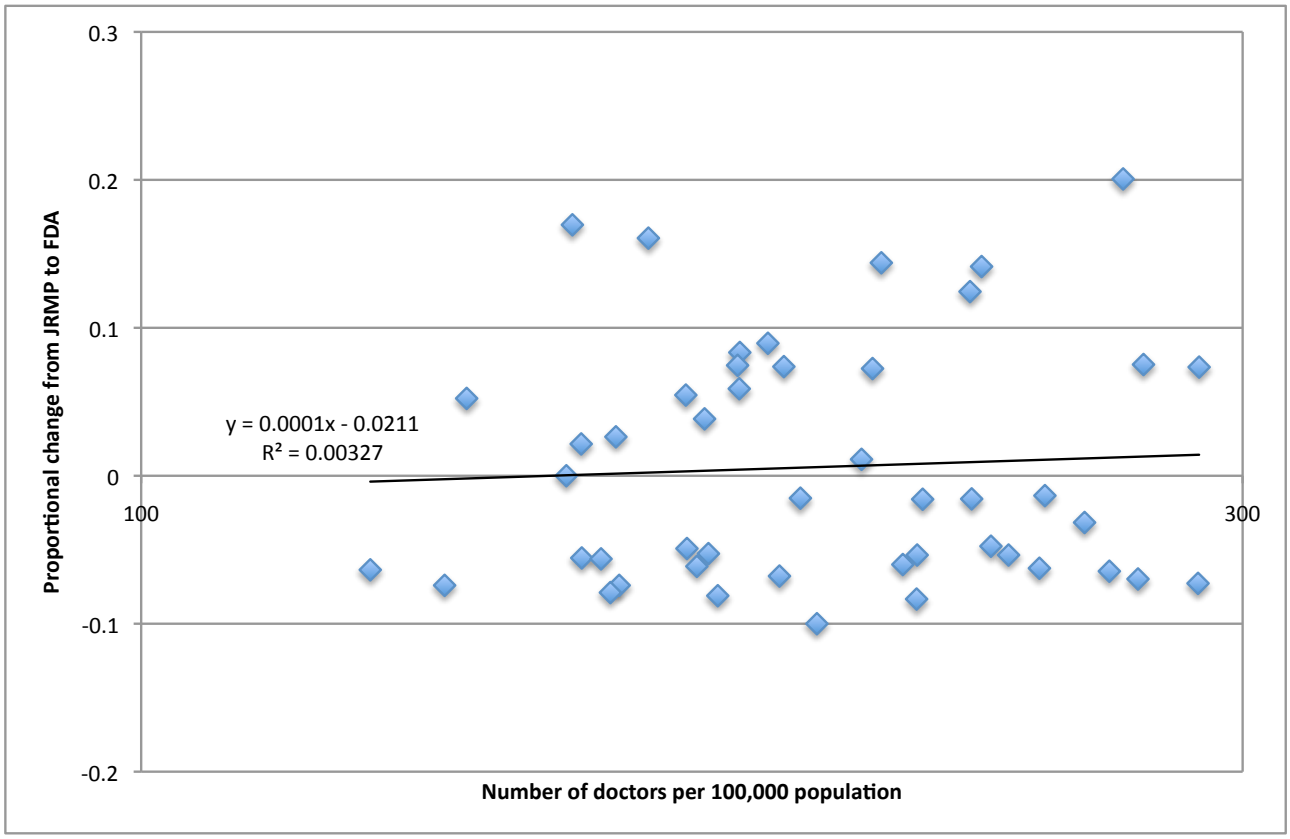

FIGURE 5. Distributional consequence of the change from JRMP to FDA.

The horizontal axis measures the number of doctors (not limited to residents) per 100,000 population from 2006, which is the latest data before 2007 that is available to us, and the vertical axis describes the proportional increase of doctor assignment in the region caused by the change of the mechanism from JRMP to FDA. The motivation for this figure is to use the number of doctors per capita as a proxy for how popular each prefecture is among doctors, and study whether there is any redistribution of doctors between popular and unpopular areas.

This figure suggests that there is virtually no adverse distributional consequence against rural areas. The linear regression suggests only a slight amount of positive relation between popularity of the prefectures and improvement/decline of doctor assignment, and indeed the R^2 value is as low as 0.00327, suggesting that there is virtually no statistical correlation between the improvement/decline of doctor assignment and how popular the area is.

Appendix C. Additional Examples

In this section we present three additional examples that present various comparative statics.

The first two examples strengthen the examples on comparative statics regarding regional preferences in the main text by showing that they hold under stronger assumptions on hospital preferences.

Example 14 (Ordering a hospital earlier may make it worse off even under homogenous hospital preferences)**.** Let there be hospitals h_1 and h_2 in region r_1, and h_3 and h_4 in region r_2. Suppose that $(q_{h_1}, q_{h_2}, q_{h_3}, q_{h_4}) = (2, 2, 2, 2)$ and $(\bar{q}_{h_1}, \bar{q}_{h_2}, \bar{q}_{h_3}, \bar{q}_{h_4}) = (1, 0, 0, 0)$. The regional cap of r_1 is 2 and that for r_2 is 1. Preferences are

$$\succ_{h_i}\colon d_1, d_2, d_3, d_4 \quad \text{for all } i = 1, \dots, 4,$$
$$\succ_{d_1}\colon h_4, h_1, \quad \succ_{d_2}\colon h_1, \quad \succ_{d_3}\colon h_2, \quad \succ_{d_4}\colon h_1, h_3.$$

We assume that h_3 is ordered earlier than h_4.

(1) Assume that h_1 is ordered earlier than h_2. In that case, in the flexible deferred acceptance mechanism, d_1 applies to h_4, d_2 and d_4 apply to h_1, and d_3 applies to h_2. d_1, d_2, and d_4 are accepted while d_3 is rejected. The matching finalizes with:

$$\mu = \begin{pmatrix} h_1 & h_2 & h_3 & h_4 & \emptyset \\ d_2, d_4 & \emptyset & \emptyset & d_1 & d_3 \end{pmatrix}.$$

(2) Assume that h_1 is ordered after h_2. In that case, in the flexible deferred acceptance mechanism, d_1 applies to h_4, d_2 and d_4 apply to h_1, and d_3 applies to h_2. d_1, d_2, and d_3 are accepted while d_4 is rejected. d_4 applies to h_3 next, and d_1 is rejected. d_1 then applies to h_1, which now rejects d_2. The matching finalizes with:

$$\mu' = \begin{pmatrix} h_1 & h_2 & h_3 & h_4 & \emptyset \\ d_1 & d_3 & d_4 & \emptyset & d_2 \end{pmatrix}.$$

First, notice that hospital h_2 is better off in case (2). Thus being ordered earlier helps h_2 in this example. However, if h_1 prefers $\{d_1\}$ to $\{d_2, d_4\}$ (which is consistent with the

assumption that hospital preferences are responsive with capacities), then h_1 is also made better off in case (2). Therefore, the effect of a picking order on hospitals' welfare is not monotone. □

Example 15 (Target monotonicity may fail even under homogenous hospital preferences)**.** Consider a market that is identical to the one in Example 8, except that the target of h_1 is now decreased to 0, with the order such that h_1 chooses before h_2. Then h_1 is matched to $\{d_1\}$ under the flexible deferred acceptance mechanism. Therefore, if h_1 prefers $\{d_1\}$ to $\{d_2, d_4\}$, then h_1 is made better off when its target capacity is smaller. □

In these examples, it is hospitals that have homogeneous preferences. However, these examples can be modified so that doctors have homogeneous preferences. To do so, modify preferences to

$$\succ_{h_1}\colon d_1, d_2, d_4, \quad \succ_{h_2}\colon d_3, \quad \succ_{h_3}\colon d_4, \quad \succ_{h_4}\colon d_1,$$
$$\succ_{d_i}\colon h_4, h_1, h_3, h_2 \quad \text{for all } i = 1, \ldots, 4.$$

That is, hospital h finds doctor d acceptable if and only if d finds h acceptable in the previous examples, while all doctors find all hospitals acceptable and the ranking between two hospitals are consistent with the rankings between two acceptable hospitals in the previous examples. By construction, the matchings produced by the flexible deferred acceptance algorithm in this market are identical to those in the previous examples.

The next example studies comparative statics. Consider splitting a region into a number of smaller regions that partition the original region, and dividing the original regional cap among the new smaller regions. One might suspect that doing so makes doctors weakly worse off because the new set of constraints based on smaller regions may appear more stringent. The following example shows that this conjecture is incorrect. In fact, splitting regions can make some doctors and hospitals strictly better off, while making other doctors and hospitals strictly worse off.

Example 16 (Splitting regions has ambiguous welfare effects)**.** Let there be three hospitals, h_i for $i = 1, 2, 3$ in the grand region r with regional cap of 1. The capacity of each hospital is 1. There are three doctors in the market, d_i for $i = 1, 2, 3$. Suppose that the regional preferences are such that $(1, 0, 0) \succ_r (0, 1, 0) \succ_r (0, 0, 1)$.

We examine the effect of splitting region r into two smaller regions, $r' = \{h_1, h_3\}$ and $r'' = \{h_2\}$. The splitting needs some rule of allocating the regional cap to the smaller regions, which in this example corresponds to allocating the cap 1 of r either to r' or to r''

(while allocating the regional cap of zero to the other region).[78] In what follows we show that in either case, there exists a preference profile such that the welfare effect of splitting is ambiguous (i.e., under such a preference profile it is not the case that every agent of one side of the market becomes weakly better/worse off) under the flexible deferred acceptance mechanism.

Suppose first that the cap 1 of r is allocated to r'. Then, suppose

$$\succ_{d_i}: h_i, \qquad \succ_{h_i}: d_i$$

for $i = 2, 3$, and d_1 and h_1 regard no one as acceptable. The flexible deferred acceptance mechanism produces a matching μ such that $\mu_{d_2} = h_2$ before splitting, while it produces a matching μ' such that $\mu'_{d_3} = h_3$ after splitting (no other doctors are matched in either matching). Thus, splitting the region r makes d_2 and h_2 strictly worse off, while making d_3 and h_3 strictly better off.

Suppose second that the cap 1 of r is allocated to r''. Then, suppose

$$\succ_{d_i}: h_i, \qquad \succ_{h_i}: d_i$$

for $i = 1, 2$, and d_3 and h_3 regard no one as acceptable. The flexible deferred acceptance mechanism produces a matching μ such that $\mu_{d_1} = h_1$ before splitting, while it produces a matching μ' such that $\mu'_{d_2} = h_2$ after splitting (no other doctors are matched in either matching). Thus, splitting the region r makes d_1 and h_1 strictly worse off, while making d_2 and h_2 strictly better off. □

Note that an analogous example can be easily constructed to show that the effect of splitting on the welfare of the hospitals outside the split region is also ambiguous. Finally, also note that the conclusion holds regardless of how we define regional preferences after splitting the grand region r.

[78]It is only for simplicity that we use an example in which the regional cap of the grand region is one, and thus one of the smaller regions has a regional cap of zero. Our conclusion does not depend on this (perhaps unrealistic) assumption: The same point can be made in examples in which a region with regional cap larger than one is split.