Does gender affect medical decisions?

Results from a behavioral experiment with physicians and medical students

GEIR GODAGER*[1,2]
HEIKE HENNIG-SCHMIDT[3,1]
JING JING LI[4,1]
JIAN WANG[5,1]
FAN YANG[1,6]

1 Institute of Health and Society, Department of Health Management and Health Economics, University of Oslo.
2 Health Services Research Unit, Akershus University Hospital.
3 BonnEconLab, University of Bonn.
4 Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, Shandong, 250021.
5 Dong Fureng Institute of Economic and Social Development, Wuhan University.
6 Department of Health Sciences in Gjøvik, Norwegian University of Science and Technology.

Abstract: It is rarely the case in medical practice that differences between female and male physicians can be described under *ceteris paribus* conditions. Physicians self-select their type of practice, patients self-select physicians, and physicians are expected to account for both the context and the characteristics of their patients when providing medical treatment. As a result, reported gender differences in medical practice can have several alternative interpretations. A key question, therefore, is whether the treatment of a *given patient* is expected to depend on the gender of the physician. To address this question, we quantify gender effects using data from an incentivized laboratory experiment, in which Chinese medical doctors and Chinese medical students choose medical treatment under different payment schemes. We estimate preference parameters of females and males assuming decision makers have patient-regarding preferences. We cannot reject the hypothesis that gender differences in treatment choices are absent. The differences between preference parameters of females and males are not statistically significant, and there is no evidence that the degree of randomness in choices differs between genders. The absence of gender effects in the laboratory, where choice context is fixed, provides nuance to previous findings on gender differences, and highlights the general difficulty of separating individuals' behavior from their context.

JEL-Classification: C92, D82, I11, H40, J33
Keywords: Gender, Laboratory experiment, Imperfect rationality, Physician behavior

1 Introduction

In this paper we ask whether the physician's gender affects the treatment patients receive. This is an important research topic, as medical care has been characterized by a substantial rise in female labor market participation worldwide due to a considerable increase of women's enrollment in medical schools and residency programs (Levinson & Lurie 2004). If female and male doctors treat their patients differently, the change in the gender composition of medical professions would change the supply of medical services and the way patients are treated. This, in particular, would be the case if patient-regarding preferences, measured by the relative weight a physician puts on patient's health benefits, differ between

*Correspondence to: Geir Godager, Department of Health Management and Health Economics, University of Oslo, Box 1089, Blindern, 0317 Oslo, Norway. E-mail: geir.godager@medisin.uio.no.
Published: Online June 2023. dx.doi.org/10.5617/njhe.10135

males and females. Knowledge on behavioral gender differences in the medical workforce is, therefore, of general and political interest. Our paper sheds light on this essential topic.

The increase of females in the medical profession has been observed in many countries, for instance, in the US, the United Kingdom, Russia, Norway, Canada, Sweden, The Netherlands, and Australia (Kilminster et al. 2007). In the United States, the proportion of female medical students rose from 10% to 50% from 1970 to 2019 (Boyle 2019), and females now constitute the majority of medical students in many countries (OECD 2020).

In Scandinavian countries, the share of female doctors amounts to about 50%, reflecting a substantial increase in recent years. In Sweden, the share of female employed doctors increased from 48% in 2018, to 53% in 2019 (Ström 2021). In Norway, the proportion of women among working doctors under the age of 70 surpassed 50% in 2017, while in 2022, 54.4% were women, with 48.4% of specialists being female (The Norwegian Medical Association 2022). In Denmark, the share of female doctors was 55% in 2021. Projections indicate a substantial rise to 66% by 2045 (Lund 2022). Female doctors have also become common in China. The share of licensed female doctors increased from 43% in 2010 to 47% in 2019, and in the hospital sector, the percentage of female doctors is even higher with 70% in 2010 and 74% in 2019 (Ministry of Health in People's Republic of China 2011, China Health Commission 2020).

This shift in the gender composition might bring about changes in medical service provision and patient treatment as has been suggested, for instance, by Riska (2001), Boulis & Jacobs (2008), and McKinstry (2008). Previous research on the gender differences in the medical profession reports mixed results, however. With regard to physician-patient communication, for instance, the evidence is inconclusive. Some studies have found that female doctors spend more time on their patients than male doctors do (Jefferson et al. 2013, Roter & Hall 2004). Others report the opposite (Hampson et al. 1996, Roter et al. 1999) while Bertakis et al. (1995) and Jefferson et al. (2015) found nearly no difference.[1] Studying gender differences in physicians' behavior by means of detailed and complete national register data on Norwegian general practitioners (GPs) has also led to inconclusive evidence. Iversen & Lurås (2002) report that female GPs offer shorter waiting time to their patients compared to male GPs. Yet, more patients decide to switch out of the patient lists of female general practitioners (Iversen & Lurås 2011). Godager et al. (2015) found no significant differences in referrals to hospitals and specialists between female and male GPs, and there was no significant difference between female and male GPs' propensity for working voluntarily for the community health service (Godager & Lurås 2009).

A key question in gender studies is whether detected behavioral differences between males and females are differences in preferences or whether they reflect the design and contextual setting of the decision environment (Niederle 2016). Finding a situation in the real world that enables such a study is nearly impossible. An economic experiment, however, can create a controlled environment in the laboratory, and thus enables the identification and quantification of gender differences in preferences and behavior. We, therefore, use data of a behavioral experiment in a stylized physician decision-making context that is stripped of many confounding factors, such as differences in patient populations, strategic behavior, discrimination or complex interactions with patients, colleagues, insurers or other third-party payers. Our design highlights the basic question in a doctor-patient interaction: How much weight does a physician put on the patient's health benefits?

[1] See also Dacre (2008) and Hedden et al. (2014), for further discussions on this topic.

Our experimental design involves individual decision-making tasks in which each female or male participant acts in the role of a physician. Participants in our experiment are Chinese medical students ($N = 178$: 101 females and 77 males). We also "brought the field to the lab" by recruiting Chinese medical doctors ($N = 99$: 69 females and 30 males) to analyze whether gender differences can occur with participants from outside the laboratory and beyond experiments with students. Our experiment applies the same experimental parameters as Hennig-Schmidt et al. (2011), and our study is based on the data of Wang et al. (2020). We study the decision-making of physicians who are consecutively confronted with two different payment mechanisms. Patients and diseases are kept constant in both payment schemes, which eliminates differences in patient populations. In our medically framed setting, the choices of the participants' determine both the physician's profit and a patient's health benefit. Decisions are incentivized by monetary rewards. Even though patients are abstract in our experiment, a real patient outside the lab is supported in that the monetary equivalent of the aggregated patient benefits is reducing the cancer treatment costs.

The mainstream economic models of decision-making typically assume *perfectly rational* actors who consistently choose alternatives that maximize their utility. As highlighted by, e.g., Kahneman & Tversky (1984), Cox (1997) and Sen (1997), the assumption of perfect rationality has little empirical support. McFadden et al. (1999) argues that the assumption of perfect rationality is unnecessarily strong given the contemporary theory and the methods available. We therefore apply empirical methods that allow for *imperfectly rational* decision-makers, while at the same time, they enable the identification of preference parameters. The structural model we use distinguishes between behavioral differences caused by (1) gender differences in the willingness to sacrifice profit in order to improve patient benefit, i.e. differences in *patient-regarding preferences*, and by (2) the propensity for choosing treatment alternatives that are inconsistent with utility-maximization, i.e. differences in the *degree of randomness in behavior*. We also analyze by parametric regressions whether males and females differ in an important consequence of randomness, namely *Pareto-dominated choices*.

One might wonder why gender differences in the degree of randomness in behavior would be expected at all. There is evidence from the psychological literature that a relationship exists between different ways of thinking (*thinking styles)*, in particular experiential and rational modes of reasoning, see Sladek et al. (2008) and the literature cited therein.[2] The study found that differences in thinking dispositions of male doctors influence their behavior with regard to following treatment guidelines. Higher scores in experiential mode of reasoning are associated with self-reported guideline discordance, while higher scores in rational thinking style are associated with overall guideline compliance. As male healthcare workers are found to score higher in rational thinking than women, and females are found to score higher in experiential reasoning compared to men (Sladek et al. 2010), behavioral differences due to differences in thinking styles and thus in the degree of randomness are not implausible.

To the best of our knowledge, our experimental study is currently the only one that explicitly analyzes the gender effect in a physician decision-making task involving doc-

[2] Experiential thinking style comprises cognitive processing characterized by holistic and emotional views on reality, and is outcome oriented. Behavior is mediated by experience. In contrast, rational thinking style comprises cognitive processing characterized by analytic and logical views on reality and is reason and process oriented. Behavior is mediated by conscious appraisal of events. See Sladek et al. (2008) for more details.

tors and medical students[3]. We are also the first to distinguish explicitly between gender differences in patient-regarding preferences, in the degree of randomness, and in Pareto-dominated choices in a medical decision-making task. As all these factors can affect treatment choice, we specify the following research questions in this paper:

I. Do females and males differ in their patient-regarding preference?

II. Do females and males differ in the degree of randomness in treatment choice?

III. Do females and males differ in making Pareto-dominated treatment choices?

For each research question, our stated null hypothesis is that a gender effect is absent. For one thing, no clear directional hypotheses regarding gender differences become apparent from the literature. Further, our pure experimental design excludes many contextual factors that could affect medical decision-making; therefore, we expect rather small differences, if any.

The latter two research questions are meaningless when perfect rationality is assumed. It is not, however, under the assumption of imperfect rationality. The reason is that the existence of a one-to-one relation between preferences and behavior observed in a given context cannot be established when decision makers are imperfectly rational. Since choices will sometimes fail to maximize the decision maker's utility, the preferences cannot be inferred from observed choices unless the degree of randomness in choice is accounted for. Our structural model analysis provides insights that reach beyond what can be achieved from a descriptive analysis of observed treatment choices. If gender differences in behavior exist, the structural models allow for distinguishing between differences relating to the willingness of sacrificing profit in order to raise the patient benefit, and differences in the degree of randomness in decision-making.

For all three research questions, we cannot reject the null-hypothesis that gender differences are absent. Thus, in our samples of Chinese medical doctors and medical students, males and females do not show statistical differences in patient-regarding preference, in the degree of randomness in decision-making, and in Pareto-dominated choices. The first result is in line with findings from the altruism-related experimental literature on gender effects, e.g., Andreoni & Vesterlund (2001), Eckel & Grossman (2008), Croson & Gneezy (2009), Niederle (2016), Boschini et al. (2018), while to date, randomness regarding gender has not been investigated.

The paper proceeds as follows: Section 2 gives a description of the experimental design and procedures, and provides a descriptive analysis of observed treatment choices over genders. Our behavioral model that allows for imperfect rationality is motivated and derived in Section 3. The results from estimating our structural model parameters are presented in Section 4. We discuss the validity of experimental results and relate these to previous experimental findings in Section 5, before concluding in Section 6. The appendix provides additional analyses and further information on experimental parameters, as well as the experimental material with which participants were provided.

[3] Brosig-Koch et al. (2020), Attema et al. (2023) and Li et al. (2022), for example, control for gender but do not make gender differences the main focus of their papers.

2 Experimental data: design, procedure and descriptives

2.1 Basics of the experimental design

Participants in our experiment act in the role of physicians, who are assumed to be concerned about their own profit π as well as about the patient benefit, B, the latter depending on the quantity of medical services $q \in 0, 1, ..., 10$. The participants' task is to choose a quantity of medical services for a given patient whose health benefit is determined by that choice[4]. Each physician decides for three different patient types with five different abstract illnesses, i.e., for 15 patients in total. The combination of patient type and illness characterizes a specific patient. Patient types differ in the health benefit they gain from the medical services. Like many theoretical papers (e.g., Ellis and McGuire, 1986; Ma, 1994; Choné and Ma, 2011) we use a concave patient benefit function that has a global optimum yielding the highest benefit to a patient. We refer to quantities smaller than the optimum as underprovision of medical care, and to quantities larger than the highest patient benefit as overprovision. The three types of patients reflect the patients' different states of health. Patients 1 to 5 of Type 1 have an intermediate state of health. Patients 6 to 10 are of Type 2 with a good state of health, and patients 11 to 15 are of Type 3 and suffer from a poor state of health.

A physician's choice of medical services simultaneously determines the patient benefit and the physician's own profit. The patient is assumed to be passive and fully insured, accepting each level of medical service provided by the physician. We apply a within-subject design in which the physician is sequentially confronted with the same 15 patients (treatment choices) in the two payment systems capitation CAP and fee-for-service FFS with either CAP in Part 1 of the experiment and FFS in Part 2, or vice versa. A feature of our design is that the order of the two payment schemes is varied to create two experimental conditions that make it possible to distinguish between the effect of payment schemes and the effect of the decision-makers accumulating experience with decision-making in the experiment (Breitmoser 2021). Under FFS, the physicians' remuneration increases in the number of medical services provided to a patient. Physicians are paid a lump sum per patient under CAP. The patient health benefit is measured in monetary terms. In our experiment, no real patients are present. However, physicians' quantity choices have consequences for a real patient outside the lab. The money corresponding to the patient benefits aggregated over all decisions was transferred to one real patient's in-hospital account to reduce his out-of-pocket payment for his cancer treatment. Thus, the participants have an incentive to care for the patient when making their decisions. We did not inform the participants about the identity of the person to whose account the money would be transferred.

Before making his or her decision, the physician gets information on her remuneration, costs and profit as well as on the patient's benefit for each quantity that can be chosen. All monetary amounts are in Token, our experimental currency, with an exchange rate of 10 Token = 1 RMB for students and 10 Token = 6 RMB for doctors (1 RMB was approximately € 0.12 at the time of the experiment).

2.2 Experimental protocol

Our experiment was conducted in September 2012 (medical students) and 2013 (doctors) at the Center for Health Economic Experiments and Public Policy at Shandong University in Jinan, China, and was programmed with z-Tree (Fischbacher 2007). All experimental mate-

[4]For the experimental parameters see Table A.1 in Appendix A1. A more detailed description of the experimental design is found in Appendix B.

rial was provided in Chinese (see Appendix C for the English version). Each of the Chinese male and female medical students and doctors participating in our experiment was sequentially confronted with the same 15 patients in FFS and CAP. The subjects were randomly assigned to experimental sessions where either CAP was implemented in Part 1 followed by FFS in Part 2 (condition CAP-FFS) or in the reversed order (condition FFS-CAP). Each participant joined the experiment only once, either in CAP-FFS or in FFS-CAP. Participants were initially informed that the experiment consisted of two parts, but were not informed about what the second part would be about. The male and female medical students, who voluntarily participated in the experiment, were recruited via notices posted at the campus and by email invitations. Doctors who were working at community health service centers in five districts of Jinan were recruited through a phone call by the respective District Department of Health informing them that a research study at Shandong University needed volunteers.

The experimental procedure was exactly the same for all medical students and doctors. Participants were randomly allocated to their workstations separated by wooden panels and curtains to guarantee anonymity of their decisions. Then, instructions for Part 1 of the experiment were distributed and read aloud by a Chinese experimenter. Subjects had ample time to read the instructions and to ask clarifying questions in private that were answered individually. Next they had to answer a set of test questions. Participants decided under either a CAP or a FFS system and went through a sequence of 15 patients with the order of patients being predetermined and kept constant across conditions. After each decision, participants were informed about their profit and the patient benefits generated by the previous choice. At the end of each part of the experiment, they received information about their total profit and the total health benefit generated during all 15 quantity decisions. At the end of Part 1, the participants answered some open-ended questions on their choice motivation. In Part 2 of the experiment, participants decided again for the same patients, but under the payment system they had not yet been confronted with. All participants answered questions on socio-demographic variables, and the doctors also stated their medical specialty and the number of years in medical practice. Finally, participants were informed about their individual total profit and the total benefit resulting from their decisions in Parts 1 and 2 of the experiment as well as on their final monetary payoff. After having been paid in private, the participants left the laboratory individually.

To ensure that the doctors and medical students trusted the experimenters to actually transfer the money derived from the patient benefit, we implemented a procedure previously used in several other experiments[5]. A monitor was randomly selected from the participants in a session. The monitor verified the amount of money corresponding to the patient benefits aggregated over all participants' decisions in the respective session. Then, the monitor and an assistant to the experimenters went by taxi to the Shandong University Cancer Hospital in Jinan and paid the corresponding amount in cash at the hospital-cashier's desk into the patient's account. We took great care to ensure that the monitor did not see the name of the real patient in order to maintain the patient's anonymity. The monitor signed a statement on the appropriate transfer of the monetary amount. After all sessions had been conducted, all participants in each session received an email informing them on the respective transfer. Each monitor in the medical student subject pool was paid an additional 50 RMB and each doctor 200 RMB.

[5]See, e.g., Hennig-Schmidt et al. (2011), Godager and Wiesen (2013), Hennig-Schmidt and Wiesen (2014), Godager et al. (2016), Brosig-Koch et al. (2016, 2017, 2020), Wang et al. (2020), Ge & Godager (2021*b*) and Ge et al. (2022)

We conducted four sessions with medical doctors, and six sessions with medical students. Each experimental session comprised one condition (CAP-FFS or FFS-CAP), and lasted for approximately 90 minutes. A female student on average earned 27 RMB (€3.20), while a male student earned 28 RMB (€3.40)plus a show-up fee of 15 RMB (€1.80). Female doctors on average earned 159 RMB (€19.10) and male doctors earned 163 RMB (€19.60) plus a show-up fee of 120 RMB (€14.40).[6] Based on all 8,310 decisions, a total of 19,814 RMB (€2,377.68) was transferred to the real patient's in-hospital-account to be used for reducing his out-of-pocket payment for cancer treatment: 4,751 RMB (€570.12) for the sessions with medical students and 15,063 RMB (€1,807,56) for the sessions with doctors.

We followed routines for ethical assessments at the University of Oslo and applied standard experimental procedures when dealing with human subjects. The experiment was conducted according to the ethics standards of the experimental economics profession that do not allow deception. Given these standards and the fact that our decision experiment did not involve any medical treatment or alike, the approval by an ethics committee or institutional review board was waved at the institution where the experiment was run. Approval of the study was given by Norwegian Social Science Data Services (reference #44267).

2.3 Descriptives

We start by analyzing the quantity choices aggregated for each of the two payment schemes. We further differentiate between male and female doctors and medical students. Table 1 provides the gender composition in the two conditions CAP-FFS and FFS-CAP and the respective frequencies of doctors and students.[7] In total, 137 subjects participated in sessions where CAP was followed by FFS, whereas 140 subjects participated in sessions where FFS was followed by CAP. We observe that participation in the two conditions is approximately balanced among both medical doctors (49 CAP-FFS, 50 FFS-CAP) and medical students (88 CAP-FFS, 90 FFS-CAP).

Table 1: Gender composition among doctors (N=99) and medical students (N=178) over the two conditions of the experiment

Condition	Gender	No of subjects (Doc.)[Stud.]
CAP-FFS	Female	76 (31)[45]
	Male	61 (18)[43]
	Total	137 (49)[88]
FFS-CAP	Female	94 (38)[56]
	Male	46 (12)[34]
	Total	140 (50)[90]

Table 2 provides the average quantity provision for each of the 15 patients under CAP and FFS by gender, with observations pooled over both parts of the experiment.

We also describe behavior at less aggregated levels. Figure 1 shows the average quantities of service provision for each of the 15 patients under both payment systems. We observe

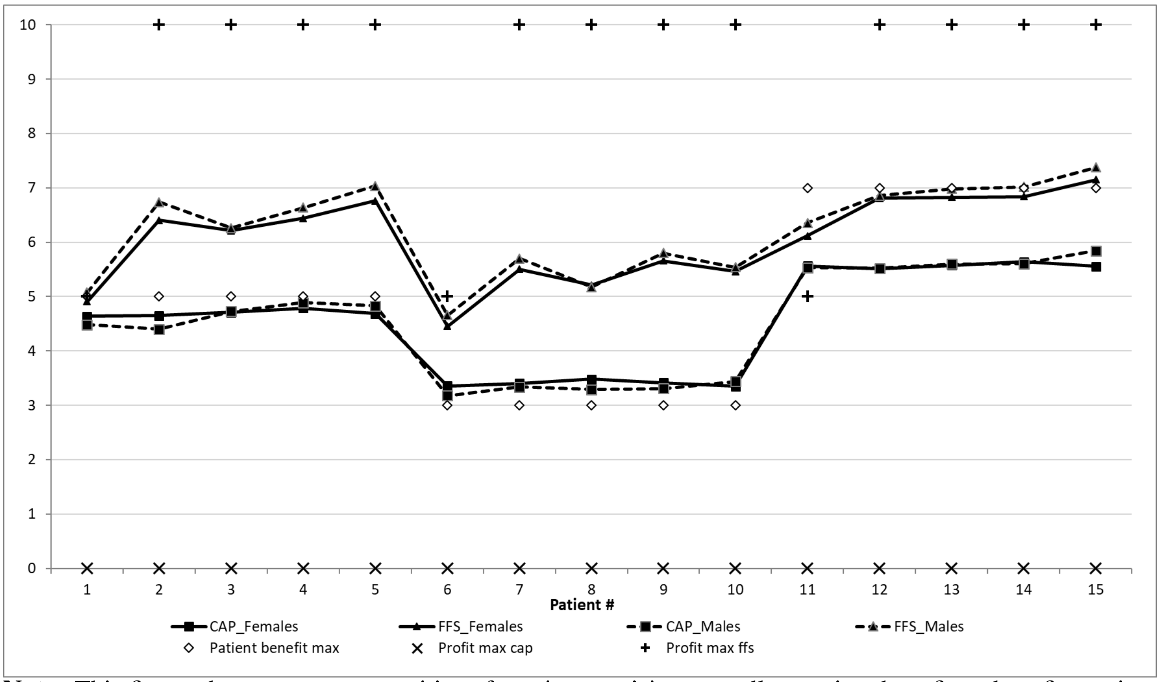
[6] We adjusted stake sizes according to opportunity costs (Herrmann et al. 2008, Gächter & Schulz 2016) outside the laboratory, i.e., the hourly wage of a student helper and the average hourly wage of a doctor in the respective employment situation. The average payoff for students approximately corresponded to the hourly wage of a student helper at Shandong University of about 30 RMB. For doctors the average hourly wage was about 120 RMB.

[7] The following analyses are based on the data set of Wang et al. (2020).

Table 2: Mean (St.dev) of chosen quantities by payment scheme and gender, for doctors and medical students

Payment	Doctors		Students	
	Female	Male	Female	Male
CAP	4.54 (1.80)	4.69 (1.73)	4.57 (1.62)	4.47 (1.51)
FFS	5.93 (2.03)	6.27 (1.63)	6.14 (1.78)	6.19 (1.79)

Figure 1: Mean quantity provision for each of the 15 patients under CAP and FFS by gender. Observations are pooled over both parts of the experiment.



Note: This figure shows average quantities of service provision as well as patient benefit and profit maxima for payment systems FFS and CAP for female/male Chinese doctors (N=99), and Chinese medical students (N=178), pooled over both parts of the experiment.

in both payment schemes that females and males behave similarly. Assuming independence of treatment choices, Mann-Whitney U-Tests (MW in the following) provide no evidence for statistical differences between females and males at a conventional significance level ($p = 0.0659$).[8] From our descriptive analysis of quantity choices we cannot reject the null hypothesis that quantity choices are unaffected by gender. Our conclusion is supported by additional reduced form analyses provided in Tables A.3 and A.4 in Appendix A3, where we report estimation results from linear and ordinal regression models with individual specific random effects.

[8]This p-value should be interpreted with caution: The experiment generates 8,310 observed quantity choices. Observations cannot be regarded as independent, since each individual subject makes 30 choices. A simple approach to acknowledging the panel structure of the data is to test for gender effects by applying MW tests repeatedly for each of the 30 treatment choices. This comes with the non-trivial challenge of appropriately adjusting the p-values to account for repeated testing. When testing the null hypothesis 30 times, the conservative Bonferroni correction (Shaffer 1995) is acquired by dividing the uncorrected p-values by 30. Since the smallest non-corrected p-value from our 30 MW-tests found is 0.0418, our null hypothesis cannot be rejected at conventional p-values. The p-value of 0.0418 was found for patient 5 in CAP. See Table A.2 in Appendix A2 for p-values for all patients.

3 Behavioral model

In this section, we motivate and derive our structural model used for quantifying gender differences in patient-regarding preferences and in the degree of randomness.

3.1 Patient-regarding preferences

The key role of patient-regarding preferences in determining physician behavior has been studied by health economists for decades (Arrow 1963, Farley 1986, Ellis & McGuire 1986, 1990, Chalkley & Malcomson 1998, Scott 2000, Jack 2005, Léger 2008, Godager et al. 2009, Iversen & Ma 2011, Chone & Ma 2011, Chandra et al. 2011, Godager et al. 2015, Liu et al. 2018, Iversen & Ma 2022). Patient-regarding preferences are typically represented by including health benefits for patients as an element in the physician's utility function. Letting n index the type of provider, B denote patient benefit, and π denote profit, providers have patient-regarding preferences when their utility function, V_n, given by:

$$V_n = V_n(\mathrm{B}, \pi) \quad , \tag{1}$$

is increasing in both B and π. Henceforth, our measure of patient-regarding preferences is the marginal rate of substitution between profit and health benefit, given by:

$$-\left.\frac{d\pi}{d\mathrm{B}}\right|_{V_n=V_n^0} = \frac{V'_{n\,1}(\mathrm{B}, \pi)}{V'_{n\,2}(\mathrm{B}, \pi)} \equiv \mathrm{MRS}_n(\mathrm{B}, \pi) \ . \tag{2}$$

$\mathrm{MRS}_n(\mathrm{B}, \pi)$ is a meaningful measure of patient-regarding preferences, sometimes referred to as the physician's willingness-to-pay for health benefits, as it represents the number of profit units the provider is willing to sacrifice to improve the patients' benefit by one unit. In the model specification of Ellis & McGuire (1986), where utility is linear in π and B, $\mathrm{MRS}_n(\mathrm{B}, \pi)$ becomes a fixed, positive parameter unaffected by π and B.[9] To address our first research question, whether males and females differ in their patient-regarding preferences, we specify a log-linear utility function given by:

$$V_n(\mathrm{B}, \pi) = \alpha_n ln(\mathrm{B}) + (1-\alpha_n) ln(\pi) \ , \quad \alpha_n \in (0,1) \ \forall n, \ n = \text{FEMALE or MALE} \tag{3}$$

with marginal rate of substitution given by:

$$\mathrm{MRS}_n(\mathrm{B}, \pi) = \frac{\alpha_n}{(1-\alpha_n)} \times \frac{\pi}{\mathrm{B}} \ . \tag{4}$$

With the assumption $\alpha_n \in (0,1)$, the utility function becomes homogeneous of degree one and exhibits constant returns to scale (CRS). The parameter α_n denotes the relative valuation of the health benefit in n's preference function. Specifying physician preferences to comprise a combination of profit and patient benefit has been shown to fit well to experimental data (Godager & Wiesen 2013, Li et al. 2017, Li 2018, Wang et al. 2020, Ge & Godager 2021*a*, Ge et al. 2022, Li et al. 2022). Since our data is from a controlled experiment guaranteeing that females and males face exactly the same values of $\frac{\pi}{\mathrm{B}}$, and since $\mathrm{MRS}_n(\mathrm{B}, \pi)$ in (4) increases monotonously with α_n, we may conclude on the question of gender differences in patient-regarding preferences by testing the null hypothesis that $\alpha_{\text{FEMALE}} = \alpha_{\text{MALE}}$.

[9]From Equation (2) we see that in case of more general specifications where second order derivatives are allowed to be different from zero, $\mathrm{MRS}_n(\mathrm{B}, \pi)$ will depend on π and B.

3.2 Randomness in behavior

Our second research question is motivated by the fact that choices violating the utility maximization assumption are frequently reported in the literature. Based on Paul Samuelson's (1938) *revealed preference principle*, Afriat (1973), Varian (1982, 1983, 1991), and others have contributed to the rich literature devoted to testing whether behaviors are actually consistent with utility maximization. As behaviors inconsistent with utility maximization were frequently found, Afriat's (1972) *"critical cost efficiency index"* and Varian's (1991 related *"violation index"* provide methods to identify a monetary value of resource waste caused by individuals failing to maximize utility. Behavior in conflict with utility maximization is generally difficult to detect, even in a favorable setting with suitable experimental data. This is because preferences can be highly heterogeneous, and observing consistency in behavior does not necessarily imply utility maximization. However, making Pareto-dominated choices results in resource waste that is observable to the researcher. Upon observing a Pareto-dominated choice, the researcher can infer that utility must be suboptimal, without needing any additional information about the individual's utility function.

In the context of our experiment, a chosen treatment alternative is Pareto-dominated if a different alternative is available that will generate (A) more profit without resulting in less patient benefit, (B) greater patient benefit without resulting in less profit, or (C) more patient benefit *and* more profit. Pareto-dominated choices are observed in 10.6% of all the decisions in our experimental data. This finding provides motivation for the empirical analysis of gender differences in making Pareto-dominated choices presented in Subsection 4.2. Moreover, this finding suggests that behavior is inconsistent with utility maximization and motivates the specification of a structural behavioral model that allows for imperfect rationality.

Structural behavioral models that account for imperfect rationality are the workhorses in the vast choice modelling literature (McFadden et al. 1999, McFadden 2001, Hess & Daly 2010, Jessie & Saari 2016, Louviere et al. 2002, Louviere & Eagle 2006). One approach to modelling imperfect rationality is to assume that a decision maker's objective is contaminated by irrelevant stochastic noise that mixes with the decision maker's utility. Behavioral models including this feature are proposed by Luce (1959), Tversky (1972) and McKelvey & Palfrey (1995). In such models, randomness in decision-making can be caused by implementation error even if the individuals' utility is constant. If, for instance , the degree of randomness in behavior varies across decision makers, they may seem heterogeneous in preferences even when they are not (Louviere & Eagle 2006). For the same reason, differences in the degree of randomness across choice contexts can make individual preferences appear context-dependent even when preferences are stable. Swait & Louviere (1993), Louviere & Eagle (2006) and Fiebig et al. (2010) argue that the degree of randomness in behavior is unlikely to be constant, as the impact of noise on choices can vary over conditions, contextual circumstances, or situations, as well as between decision makers. For example, if subjects are acquiring experience during the course of a laboratory experiment, and the researcher applies an empirical strategy that (silently) assumes the degree of randomness to be constant, the researcher might erroneously conclude that preferences changed during the experiment.

We assume that individuals' objectives are contaminated by irrelevant factors, represented by the inclusion of an unobservable noise term, ϵ_n, in the objective function, F_n:

$$F_n = V_n(\mathrm{B}, \pi) + \sigma_n \epsilon_n \; , \; \sigma_n > 0 \; \forall n, \; n = \text{FEMALE or MALE} \; . \tag{5}$$

Rather than assuming perfectly rational providers who maximize V_n in (1), we assume that providers are imperfectly rational and maximize F_n in (5). The specification in (5) reflects the assumption that rationality is present to some degree. The first term is the rational part of the individual's objective, which is the individual's deterministic utility (Train 2009) as a function of health benefits and profit. The second term comprises factors that are irrelevant for utility, and these irrelevant factors are represented by an error component, ϵ_n, weighted by σ_n. The σ_n parameter is a measure of the degree in which the individual's objective is affected by these irrelevant factors. In the corner solution where $\sigma_n = 0$, the behavior of an imperfectly rational individual maximizing F_n in Equation (5) will obviously coincide with the behavior of a perfectly rational individual maximizing $V_n(\text{B}, \pi)$ in Equation (1). Since our data come from an experiment with 30 decisions per individual, there is sufficient within-individual variation in the data to identify σ_n for individuals or groups of participants under the assumption that the utility function exhibits CRS.

To address our second research question, whether males and females differ in the degree of randomness in behavior, we quantify σ_n separately for females and males. We test for gender-differences in randomness by testing the null hypothesis $\sigma_{\text{FEMALE}} = \sigma_{\text{MALE}}$. To address our third research question on whether gender-differences exist in making Pareto-dominated choices, we test the null hypothesis that both genders have the same probability of choosing Pareto-dominated alternatives.

4 Estimation and results

In this section, we estimate the parameters of our structural model and derive our results on gender differences in patient-regarding preferences, the degree of randomness, and regarding Pareto-dominated choices (Research Questions I, II, and III).

4.1 Gender differences in patient-regarding preferences and in the degree of randomness

Our empirical specification builds on the early work of Luce (1959), Tversky (1972) and McFadden (1974), as well as on the more recent literature on explicitly scaled choice models (Swait & Louviere 1993, Hole et al. 2006, Fiebig et al. 2010, Bech et al. 2011, Hess & Rose 2012, Swait & Marley 2013, Hess & Train 2017, Wallin et al. 2018, Wang et al. 2020, Ge & Godager 2021*a*). The conventional way of deriving a choice model as described by Train (2009), is to assume individuals who maximize random utility, and let random utility be the sum of a deterministic utility term and a random term. As highlighted by Hess & Rose (2012) and Hess & Train (2017), the model we apply in this paper is in practice the same as traditional textbook models. The motivation and interpretation differ, however. Similar to Luce (1959) and McKelvey & Palfrey (1995), we highlight imperfect rationality as a source of randomness in behavior rather than explaining randomness in behavior as driven by factors that are unobservable to the researcher, as in McFadden (1974).[10]

Our empirical specification given in Equation (6) is achieved by inserting (3) into (5),

[10]As described by Ge & Godager (2021*a*, Section 2), there are several sources of observed randomness in human behavior.

introducing necessary indexes and specifying the structure of the errors ϵ_n:

$$F_{njt} = \alpha_n ln(\mathrm{B}_{jt}) + (1-\alpha_n) ln(\pi_{jt}) + \sigma_{nt}[a_j + \varepsilon_{njt}] \quad \alpha_n \in (0,1)\ \forall n \ , \tag{6}$$
$$\sigma_{nt} = \mathrm{e}^{(\boldsymbol{\theta} \times \boldsymbol{z})} \ . \tag{7}$$

Equation (6) implies a CRS assumption introduced by the requirement that the preference parameters assigned to patient benefit and profit sum to one for each decision maker. The CRS assumption introduces constraints that enable the identification of σ_{nt}. Our model is, therefore, a *scaled logit model* (Fiebig et al. 2010). In a scaled logit model, identification of model parameters is achieved by implementing constraints on attribute coefficients instead of normalizing the variance of the error terms as in a conventional *conditional logit model*. The CRS approach introduces a set of constraints by assuming the two preference parameters to sum to one for each n. It is the addition of these restrictions that enables us to estimate the so-called *scale parameters* σ_{nt}.[11] Equation (7) specifies the σ_{nt} as a function of observable variables.

The convention in the choice modelling literature is to distinguish between the terms *choice occasion*, which refers to a particular decision-making situation, *choice alternative*, which refers to the alternatives available to a decision maker in a choice occasion, and the *choice*, which refers to a particular alternative chosen in a particular choice occasion. We use the index t for the 30 occasions where treatment choices are made, 15 in each of the payment schemes CAP or FFS in Part 1 or 2, respectively. The index n denotes the decision maker group, interpreted here as female or male, medical student or medical doctor. We use j to index the eleven different treatment alternatives (quantities of service provision, $q \in 0, 1, ..., 10$) that are available for each patient. Our model includes alternative specific constants, denoted by a_j. Similar to Fiebig et al. (2010), a_j-terms are assumed to be part of the error structure. The error component in our model is thus given by: $\epsilon_n = a_j + \varepsilon_{njt}$. In textbook applications, the ε_{njt} terms are commonly assumed to be independent, type 1 extreme value distributed (Train 2009), and this is a sufficient, but not a necessary, condition for ensuring that choice probabilities are given by the logit formula.[12]

Equation (7) specifies the strictly positive scale parameter as a function of a vector of observable variables, $\boldsymbol{z}$. This vector includes the dummy variable DOCTOR, which is equal to one if the decision maker is a doctor, the dummy variable EXPERIENCE, which is equal to one for decisions in Part 2 of the experiment, and the dummy variable FEMALE equal to one if the decision maker is a female. As in Wang et al. (2020), the $\boldsymbol{z}$-vector also includes dummies for choice occasion. By including these dummies, we account for the scenario that the impact of noise on treatment choice depends on the decision-making task.

We use the program gmnl in STATA 17, written by Gu et al. (2013) to estimate the

[11]We follow the terminology and notation in Train (2009), and refer to σ_{nt} as the scale parameter. There is a difference in notation between Train (2009) and Fiebig et al. (2010). Train (2009, p 40-41) refers to the σ as the *scale parameter*, while Fiebig et al. (2010) refer to σ^{-1} as the *scale of the error term* on page 397, right column, which corresponds to the *rationality parameter* λ in McKelvey & Palfrey (1995). We are not the first to apply the CRS assumption as an identification strategy. See, e.g., Swait & Marley (2013), Wallin et al. (2018), Wang et al. (2020) and Ge & Godager (2021*a*). A third identification strategy is to specify so-called *willingness-to-pay space models* as in the studies by Train & Weeks (2005), Scarpa et al. (2008) and Hole & Kolstad (2012).

[12]Choice probabilities given by the logit formula can be derived axiomatically under weaker assumptions (Dagsvik 1995, Erlander 1998, Dagsvik 2008, 2016, 2018), hence the specification of logit models to characterize human choices does not rely on strong assumptions.

parameters of our behavioral model (6) by means of maximum likelihood. In Table 3, we report the results. The point estimates of α reported in Panel I are 0.46 for females and 0.49 for males, and their estimated confidence intervals overlap. We cannot reject the hypothesis that $\alpha_{\mathrm{F}} = \alpha_{\mathrm{M}}$ (p-value=0.6375, Wald test).

Table 3: Results from maximum likelihood estimation of the model in Equation (6)

	Estimate	(95% C.I.)	P-value
I. **Preference parameters**			
α_{F}	0.46	(0.33— 0.60)	< 0.001
α_{M}	0.49	(0.34— 0.64)	< 0.001
II. **Scale heterogeneity**†			
θ_{DOCTOR}	0.58	(0.15— 1.00)	0.008
$\theta_{\mathrm{EXPERIENCE}}$	-0.49	(-0.80— -0.18)	0.002
θ_{FEMALE}	0.26	(-0.07— 0.58)	0.125

Note: The sample includes 178 Chinese students and 99 Chinese doctors, 30 decisions for each Chinese subject. EXPERIENCE=1 when subjects decide in Part 2 of the experimental session. P-values and C.I. are based on standard errors clustered at the level of each individual. Alternative- and patient-specific constants not shown.† The θ parameters we report are obtained by multiplying with -1 the θ parameters that are provided by the program of Gu et al. (2013). θ-parameters are marginal effects on the log of the scale parameter: $\frac{\delta ln(\sigma)}{\delta z}$.

The estimated preference parameters have clear economic interpretation. Our log-linear specification implies that the relative willingness-to-pay (RWTP) is a constant given by:

$$- \frac{d\pi}{dB} \frac{\mathrm{B}}{\pi}\bigg|_{V=V^0} \times \frac{\mathrm{B}}{\pi} = \frac{\alpha_n}{1-\alpha_n} \quad . \tag{8}$$

The RWTP in (8) is the percentage sacrifice in profit that will render the decision maker's utility unchanged if the patient benefit is increased by one percent.[13] Using the formula in (8), we find that the point estimates of the two RWTPs are 0.86 for females and 0.97 for males, meaning that females (males) are willing to forgo 0.86 (0.97) percent of their profit to raise the patient benefit by one percent.

We also estimated a fully flexible model where each of the four groups, female doctors, male doctors, female students and male students had group-specific α and σ parameters. We compared this fully flexible model to a restricted model where preference parameters were constrained to be identical for the four groups. We could not reject the null hypothesis that the most flexible model does not provide a better fit to the data than the restricted model (p-value= 0.3109, likelihood-ratio test). We conclude that the parsimonious model in Table 3 is sufficient for addressing Research Questions I and II.

With reference to Research Question I, we state:

RESULT I: We do not find a gender difference in patient-regarding preferences.

The θ_{FEMALE}-parameter is our measure of gender differences in the *degree of randomness*. We observe in Panel II of Table 3 that θ_{FEMALE} is not statistically significant. We cannot reject the hypothesis that females and males are equally influenced by irrelevant aspects when choosing medical treatments.

[13]The RWTP should not be confused with the *elasticity of substitution*, which in case of the Cobb-Douglas function with constant return to scale, is a given constant equal to one.

With reference to Research Question II, we state:

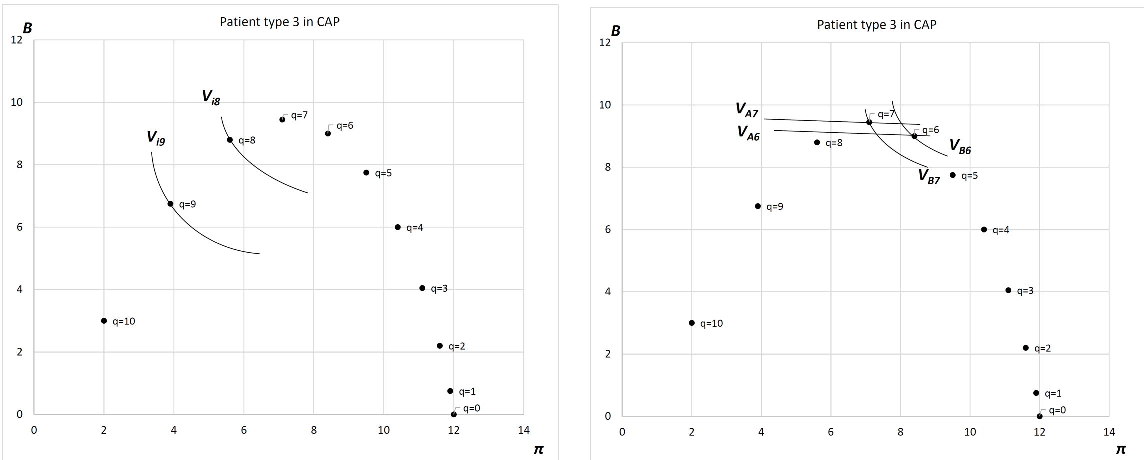
> **RESULT II:** We do not find a gender difference in the degree of randomness in treatment choices.

Panel II in Table 3 also provides evidence of significant differences in the degree of randomness in behavior. First, medical doctors are more random in their choices than medical students. Further, randomness decreases significantly when subjects are experienced, i.e., in Part 2 of the experiment. While the degree of randomness in treatment choices does not differ significantly between males and females, the estimated gender difference is smaller than the difference caused by having experienced 15 treatment decisions in Part 1 of the experiment.[14]

4.2 Gender differences in Pareto-dominated choices

In this subsection, we address Research Question III concerning the existence of gender differences in making Pareto-dominated treatment choices, which create observable resource waste inconsistent with utility maximization. By definition, a given treatment alternative is Pareto-dominated if a different alternative is available that will generate (A) more profit without resulting in less patient benefit, (B) more patient benefit without resulting in less profit, or (C) more patient benefit *and* more profit. This analysis complements the analysis of randomness in behavior by analyzing a specific type of observable behavior that can be categorized as being inconsistent with utility maximization.

Figure 2: Example of a Pareto-dominated treatment choice for patient type 3 in CAP. In this example, choosing $q = 7$ results in more profit and more patient benefit compared to each of the alternatives $q = 8$, $q = 9$ and $q = 10$. Alternatives with $q > 7$ are, therefore, considered Pareto-dominated.



a) An individual i assumed to derive utility from profit π and from patient benefit B, cannot be a perfectly rational utility maximizer when providing, for example, $q = 9$ for patient type 3 in CAP. Here, individual i would create more patient benefit *and* more profit if $q = 8$ is chosen instead of $q = 9$. For all individuals who derive utility from both π and B, utility from providing $q = 8$ (V_{i8}) must exceed utility from providing $q = 9$ (V_{i9})

b) For $q < 7$, the individual must sacrifice profit in order to increase the benefit for patients. Hence, there are no Pareto-dominated alternatives to the right of $q = 7$, where $q < 7$. We see that individual A prefers $q = 7$ over $q = 6$ since indifference curve V_{A7} is further from the origin than V_{A6}. Individual B, who is less altruistic, prefers $q = 6$ over $q = 7$ since indifference curve V_{B6} is further from the origin than V_{B7}.

[14] See Appendix A3. for more details on the behavioral differences in the degree of randomness.

We start by giving examples of Pareto-dominated treatment alternatives in our experiment. Panel (a) of Figure 2 illustrates why the set of treatment alternatives such that $q > 7$ for patient type 3 in CAP are Pareto-dominated alternatives, and provides examples of choices inconsistent with maximizing the increasing function $V_n(\text{B}, \pi)$: Choosing $q > 7$ must yield strictly lower utility than choosing $q = 7$, regardless of whether the provider is a profit maximizer ($\alpha_n = 0$), a patient benefit maximizer ($\alpha_n = 1$), or partially altruistic with $\alpha_n \in (0, 1)$. Panel (b) of Figure 2 shows that the set of treatment alternatives such that $q \leq 7$ for patient type 3 in CAP does not include any Pareto-dominated alternatives. For $q < 7$, more patient benefit can only be obtained by sacrificing profit, which implies that none of the alternatives that constitute underprovision are Pareto-dominated.

The percentage of Pareto-dominated choices made by males and females in the experiment is reported in Table 4. To test for gender-differences in the probability of making Pareto-dominated choices while accounting for repeated observations by the same decision maker, we estimate a logit model with random effects. Results from maximum likelihood estimation of this logit model are reported in Table 5. The estimated gender coefficient is not statistically significant. We cannot reject the null hypothesis that the probability of choosing Pareto-dominated treatment alternatives is the same for both genders.

Table 4: Percentage of Pareto-dominated choices by gender, experience and payment scheme

	CAP		FFS	
	PART 1: EXPERIENCE=0	PART 2: EXPERIENCE=1	PART 1: EXPERIENCE=0	PART 2: EXPERIENCE=1
FEMALES	15.8 %	6.0 %	13.5 %	11.5 %
MALES	13.1 %	1.7 %	11.9 %	8.9 %

Note: This table shows the percentage of Pareto-dominated choices by gender, experience and payment scheme. For females' decisions under cap, EXPERIENCE=0 applies for 1,140 decisions, EXPERIENCE=1 applies for 1,410 decisions. For females' decisions under FFS, EXPERIENCE=0 applies for 1,410 decisions, EXPERIENCE=1 applies for 1,140 decisions. For males' decisions under CAP, EXPERIENCE=0 applies for 915 decisions, EXPERIENCE=1 applies for 690 decisions. For males' decisions under FFS, EXPERIENCE=0 applies for 690 decisions, EXPERIENCE=1 applies for 915 decisions.

Table 5: Pareto-dominated choices by gender, subject pool, and experience. Estimation results from a logit regression with random effects. N=277 subjects, T=30 treatment choices.

Dependent variable: Dummy variable=1 if a pareto-dominated quantity was chosen		
	Estimate	Robust Std. Err.†
FEMALE	0.08	0.19
DOCTOR	1.16	0.18***
EXPERIENCE	-0.85	0.15***
FFS	0.37	0.15
constant	-3.04	0.19 ***
rho = 0.31	(fraction of variance due to u_i)	

† Clustered at the level of the individual decision-maker.***(**)[*] indicate statistically significant parameter with p-value <0.0001(<0.001)[<0.01] in a two-sided test.

Table 5 provides some additional insights. Pareto-dominated treatment choices are more frequent when participants decide in Part 1 of the experiment, i.e., when they are inexperienced with the decision situation EXPERIENCE=0, compared to Part 2 when they have had some previous experience EXPERIENCE=1, and this difference is significant. As shown in Table 5, Chinese doctors make significantly more Pareto-dominated treatment choices than Chinese students.

With reference to Research Question III, we state:

> **RESULT III:** We do not find a gender difference in the probability of choosing Pareto-dominated treatment choices.

5 Discussion

We are not the first to apply controlled experiments for studying gender differences in behavior. For instance, non-medical experiments on the dictator game, which is related to our design, (Forsythe et al. 1994, Cappelen et al. 2007, 2013, Almås et al. 2020), and on donations to a charity (Eckel & Grossman 1996, Grossman & Eckel 2015) are frequently applied tools when studying other-regarding preferences. The surveys by Eckel & Grossman (2008), Croson & Gneezy (2009) and Niederle (2016) report mixed evidence concerning gender differences. Regarding medical decision-making, our experimental study, to the best of our knowledge, is currently the only one that explicitly analyzes the gender effect, is framed in a medical context making the needs of a real patient salient, and involves doctors and medical students. We are also the first to distinguish between gender differences in patient-regarding preferences and in the degree of randomness.

The studies by Li et al. (2017), Li (2018) and Li et al. (2022)) have some similarity with our design, as they are based on a dictator game. The dictator (medical students or doctors) simultaneously determines own profit and the amount a recipient gets, and the latter amount can be increased if the dictator sacrifices own profit. Important differences to our study are that the game is neutrally framed, the medical context is not made salient to the participants (the notion of physician and patient is absent) and recipients are not patients. Li and her coauthors find no gender differences in altruism.

Many features of our experiment have been carefully chosen to add relevance. For one thing, our protocol employs a medical context framing in order to make participants feel familiar with their professional decision-making. Second, the ethical norm of altruistic – in our scenario: patient-regarding – behavior, was made salient (Eckel & Grossman 1996, Grossman & Eckel 2015) by transferring the monetary equivalent of the aggregated patient benefits to a real patient in need of financial support for the expensive medical treatment to survive cancer. Third, our participant sample consists of decision makers who are or will be real actors in the field. The doctors had approximately 16 years of professional experience on average. The prospective physicians were not newcomers either, as their average duration of medical study was about five semesters. Finally, many participants stressed the relevance of our design features in relation to their professional decision-making environment in the open questions about the factors that influenced their decisions. Thus, our design may induce the ethical standards of a medical professional to become relatively strong in the experiment compared to the field setting, thereby contributing to the observed absence of gender differences in patient-regarding behavior.

Our results are obtained in a stylized physician decision-making context stripped of contextual factors that may confound with the patient-regarding attitude of the physician or the medical student decision maker. The situation is thus, as much as possible, reduced to a very basic doctor-patient relationship. This strength of providing a controlled but artificial situation that facilitates causal inference, at the same time may be seen as the weakness of the experimental method. The decision context in the laboratory obviously differs from that of a real doctor-patient encounter. We are, however, convinced that analyzing behavior also in a laboratory setting is important, as such behavioral studies provide an additional piece of evidence, a broader view on and a better understanding of gender differences in general and in medical decision-making in particular. Such studies are nearly impossible in the field and are, therefore, complementary to field studies.

Our results suggest that patient-regarding preferences of females and males are similar and that treatment choices of females and males are affected by irrelevant factors to a similar degree. One may argue that the failure to reject our null hypotheses could be the result of insufficient statistical power. Given the moderately sized samples of females and males, we cannot rule out the possibility that small gender differences in patient-regarding preferences exist. We note, however, that the gender difference in point estimates of the α parameters is relatively small. We also note that for the degree of randomness in behavior, the difference in randomness generated by having 15 decisions in each stage of experience is about two times larger than the estimated gender difference.

We use data from an experiment that was not designed for the specific purpose of analyzing gender differences in decision-making. However, the experiment of Wang et al. (2020) comprises several design features that make the data set suitable for identifying behavioral differences between males and females. Including a relatively large number of different treatment choices and varying the order of the two payment schemes introduces circular permutation of decision-making tasks. The design thus provides us with experimental data with rich and controlled variation in trade-offs between profit and patient benefit, and in experience of the decision maker. Controlling for what Breitmoser (2021) refers to as *presentation effects* - in our case whether a particular set of tasks is performed in Part 1 or Part 2 of the experiment - is necessary for avoiding omitted variable bias in estimated preference parameters. Since both genders face identical sets of decision tasks, the experiment enables the identification of gender-specific preference parameters, and of differences in the degree of randomness in behavior between females and males. The protocol of Wang et al. (2020) also allows us to identify experience effects that occur in the laboratory experiment. If, for example, all participants of the experiment had decided under CAP before FFS, deciding under FFS would be collinear with being experienced and, as a result, would make it impossible to separate experience effects from the effect of performing a different task.

6 Conclusion

In this paper we investigate whether females and males differ in their choices of medical treatment, in their patient-regarding preferences, in their degree of randomness in behavior, and in Pareto-dominated treatment choices. The research questions are motivated by the fact that the share of females employed in the health care sector has risen sharply over recent decades, and, if gender differences exist, they might bring about changes in the provision of medical care. We apply data from a fully incentivized laboratory experiment (Wang et al. 2020) based on the experimental design of (Hennig-Schmidt et al. 2011). Our use of data from a controlled laboratory experiment enables the identification of gender differences holding decision context fixed. We analyze the data by means of non-parametric and parametric methods. Based on non-parametric tests, we do not find evidence that gender affects treatment choices. We estimate a scaled choice model to test whether patient-regarding preferences or the degree of randomness in treatment choices differ between females and males. Our measure of patient-regarding preferences is the decision maker's willingness to sacrifice profit in order to raise the patient benefit. We do not find evidence that patient-regarding preferences differ over genders, nor do we find evidence that one gender behaves more randomly than the other.

While much research has aimed to shed light on the causal mechanisms behind observed gender differences, there is no unanimity in the scientific conclusions. Gender differences in education attainment and labor market participation have changed remarkably over time (Goldin et al. 2006). This suggests that understanding the contexts in which individuals make economic decisions is important when aiming at providing new knowledge on the causes of observed gender differences in behavior. This is particularly important when providers of medical services are involved. As highlighted by Niederle (2016), a key question is whether detected differences between males and females reflect the design and contextual setting of the decision environment or whether they point to innate differences in preferences and non-rational behavior. Our study provides the example of a scenario in which many confounding factors are eliminated, and where we are unable to find a gender difference in choices of medical treatment. We are aware that not finding gender differences in our stylized laboratory context does not preclude their existence in some real-world scenario. This points to the need for additional future research to address whether, and if so how, institutional contexts influence gender differences in medical service provision.

Acknowledgements

Comments, questions and suggestions from two anonymous reviewers have contributed to substantial improvements of this paper. Funding from the Research Council of Norway through the Norwegian Centre for Health Services Research (NORCHER project number 296114) and the International Research Collaborative for Health Economic Experiments (IRECOHEX project number 231776) is gratefully acknowledged.

References

Afriat, S. N. (1972), 'Efficiency estimation of production functions', *International Economic Review* **13**, 568–598.

Afriat, S. N. (1973), 'On a system of inequalities in demand analysis: an extension of the classical method', *International Economic Review* **14**, 460–472.

Almås, I., Cappelen, A. W. & Tungodden, B. (2020), 'Cutthroat capitalism versus cuddly socialism: Are americans more meritocratic and efficiency-seeking than scandinavians?', *Journal of Political Economy* **128**(5), 1753–1788.

Andreoni, J. & Vesterlund, L. (2001), 'Which is the fair sex? gender differences in altruism', *The Quarterly Journal of Economics* **116**(1), 293–312.

Arrow, K. J. (1963), 'Uncertainty and the Welfare Economics of Medical Care', *American Economic Review* **53**, 941–969.

Attema, A. E., Galizzi, M. M., Groß, M., Hennig-Schmidt, H., Karay, Y., L'Haridon, O. & Wiesen, D. (2023), 'The formation of physician altruism', *Journal of Health Economics* **87**, 102716.

Bech, M., Kjær, T. & Lauridsen, J. (2011), 'Does the number of choice sets matter? results from a web survey applying a discrete choice experiment', *Health Economics* **20**(3), 273–286.

Bertakis, K. D., Helms, L. J., Callahan, E. J., Azari, R. & Robbins, J. A. (1995), 'The influence of gender on physician practice style.', *Medical Care* .

Boschini, A., Dreber, A., von Essen, E., Muren, A. & Ranehill, E. (2018), 'Gender and altruism in a random sample', *Journal of Behavioral and Experimental Economics* **77**, 72 – 77.

Boulis, A. K. & Jacobs, J. A. (2008), *The changing face of medicine: women doctors and the evolution of health care in America*, The culture and politics of health care work, ILR Press.

Boyle, P. (2019), 'More women than men are enrolled in medical school'. (Accessed July 20 2022.). **URL:** *https://www.aamc.org/news-insights/more-women-men-are-enrolled-medical-school*

Breitmoser, Y. (2021), 'Controlling for presentation effects in choice', *Quantitative Economics* **12**(1), 251–281.

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N. & Wiesen, D. (2016), 'Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision', *Journal of Economic Behavior & Organization* **131**, 17–23.

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N. & Wiesen, D. (2017), 'The effects of introducing mixed payment systems for physicians: Experimental evidence', *Health Economics* **26**(2), 243–262.

Brosig-Koch, J., Hennig-Schmidt, H., Kokot, J., Kairies-Schwarz, N. & Wiesen, D. (2020), Physician performance pay: Experimental evidence., Technical report, University of Oslo, Health Economics Research Network.

Cappelen, A., Hole, A. D., Sørensen, E. Ø. & Tungodden, B. (2007), 'The pluralism of fairness ideals: An experimental approach', *American Economic Review* **97**(3), 818.

Cappelen, A., Konow, J., Sørensen, E. Ø. & Tungodden, B. (2013), 'Just luck: An experimental test of risk taking and fairness', *American Economic Review* **103**(4), 1398–1413.

Chalkley, M. & Malcomson, J. M. (1998), 'Contracting for health services when patient demand does not reflect quality', *Journal of Health Economics* **17**(1), 1–19.

Chandra, A., Cutler, D. & Song, Z. (2011), Who ordered that? the economics of treatment choices in medical care, *in* M. V. Pauly, T. G. Mcguire & P. P. Barros, eds, 'Handbook of Health Economics', Vol. 2, Elsevier, pp. 397–432.

China Health Commission, ed. (2020), *China Health Statistic Yearbook*, Peking Union Medical College Press, Beijing.

Chone, P. & Ma, C. A. (2011), 'Optimal Health Care Contract under Physician Agency', *Annales d'Economie et de Statistique* **101/202**, 229–256.

Cox, J. C. (1997), 'On testing the utility hypothesis', *The Economic Journal* **107**(443), 1054–1078.

Croson, R. & Gneezy, U. (2009), 'Gender differences in preferences', *Journal of Economic literature* **47**(2), 448–74.

Dacre, J. (2008), 'Are there too many female medical graduates? no', *BMJ* **336**(7647), 749–749.

Dagsvik, J. K. (1995), 'How large is the class of generalized extreme value random utility models?', *Journal of Mathematical Psychology* **39**(1), 90–98.

Dagsvik, J. K. (2008), 'Axiomatization of stochastic models for choice under uncertainty', *Mathematical Social Sciences* **55**(3), 341–370.

Dagsvik, J. K. (2016), 'What independent random utility representations are equivalent to the IIA assumption?', *Theory and Decision* **80**(3), 495–499.

Dagsvik, J. K. (2018), 'Invariance axioms and functional form restrictions in structural models', *Mathematical Social Sciences* **91**, 85–95.

Eckel, C. C. & Grossman, P. J. (2008), Differences in the economic decisions of men and women: Experimental evidence, *in* C. R. Plott & V. L. Smith, eds, 'Handbook of Experimental Economics Results, Chapter 57', Vol. 1, Elsevier, pp. 509 – 519.

Eckel, C. & Grossman, P. (1996), 'Altruism in Anonymous Dictator Games', *Games and Economic Behavior* **16**, 181–191.

Ellis, R. P. & McGuire, T. G. (1986), 'Provider Behavior under Prospective Reimbursement: Cost Sharing and Supply', *Journal of Health Economics* **5**, 129–151.

Ellis, R. P. & McGuire, T. G. (1990), 'Optimal payment systems for health services', *Journal of Health Economics* **9**, 375–396.

Erlander, S. (1998), 'Efficiency and the logit model', *Annals of Operations Research* **82**, 203–218.

Farley, P. J. (1986), 'Theories of the price and quantity of physician services: A synthesis and critique', *Journal of Health Economics* **5**, 315–333.

Fiebig, D. G., Keane, M. P., Louviere, J. & Wasi, N. (2010), 'The generalized multinomial logit model: accounting for scale and coefficient heterogeneity', *Marketing Science* **29**(3), 393–421.

Fischbacher, U. (2007), 'Z-tree: Zurich Toolboox for Readymade Economic Experiments – Experimenter's Manual', *Experimental Economics* **10**, 171–178.

Forsythe, R., Horowitz, J. L., Savin, N. E. & Sefton, M. (1994), 'Fairness in simple bargaining experiments', *Games and Economic Behavior* **6**(3), 347–369.

Gächter, S. & Schulz, J. F. (2016), 'Intrinsic honesty and the prevalence of rule violations across societies', *Nature* **531**(7595), 496.

Ge, G. & Godager, G. (2021*a*), 'Predicting strategic medical choices: An application of a quantal response equilibrium choice model', *Journal of Choice Modelling* **39**.

Ge, G. & Godager, G. (2021*b*), 'Data from an incentivized laboratory experiment on strategic medical choices', *Data in Brief* **35**, 106926.

Ge, G., Godager, G. & Wang, J. (2022), 'Exploring physician agency under demand-side cost sharing—an experimental approach', *Health Economics* **31**(6), 1202–1227.

Godager, G., Hennig-Schmidt, H. & Iversen, T. (2016), 'Does performance disclosure influence physicians' medical decisions? an experimental study', *Journal of Economic Behavior & Organization* **131**, 36–46.

Godager, G., Iversen, T. & Ma, C. A. (2009), 'Service motives and profit incentives among physicians', *International journal of health care finance and economics* **9**(1), 39–57.

Godager, G., Iversen, T. & Ma, C. A. (2015), 'Competition, gatekeeping, and health care access', *Journal of Health Economics* **39**, 159–170.

Godager, G. & Lurås, H. (2009), 'Dual job holding general practitioners: the effect of patient shortage', *Health Economics* **18**(10), 1133–1145.

Godager, G. & Wiesen, D. (2013), 'Profit or patients' health benefit? exploring the heterogeneity in physician altruism', *Journal of Health Economics* **32**, 1105–116.

Goldin, C., Katz, L. F. & Kuziemko, I. (2006), 'The homecoming of american college women: The reversal of the college gender gap', *Journal of Economic Perspectives* **20**(4), 133–156.

Grossman, P. J. & Eckel, C. C. (2015), 'Giving versus taking for a cause', *Economics Letters* **132**, 28–30.

Gu, Y., Hole, A. R. & Knox, S. (2013), 'Fitting the generalized multinomial logit model in stata', *Stata Journal* **13**(2), 382–397.

Hampson, S. E., McKay, H. G. & Glasgow, R. E. (1996), 'Patient-physician interactions in diabetes management: consistencies and variation in the structure and content of two consultations', *Patient Education and Counseling* **29**(1), 49–58.

Hedden, L., Barer, M. L., Cardiff, K., McGrail, K. M., Law, M. R. & Bourgeault, I. L. (2014), 'The implications of the feminization of the primary care physician workforce on service supply: a systematic review', *Human Resources for Health* **12**(1), 32.

Hennig-Schmidt, H., Selten, R. & Wiesen, D. (2011), 'How Payment Systems Affect Physicians' Provision Behavior – An Experimental Investigation', *Journal of Health Economics* **30**, 637–646.

Hennig-Schmidt, H. & Wiesen, D. (2014), 'Other-regarding behavior and motivation in health care provision: An experiment with medical and non-medical students', *Social Science & Medicine* **108**, 156–165.

Herrmann, B., Thöni, C. & Gächter, S. (2008), 'Antisocial punishment across societies', *Science* **319**(5868), 1362–1367.

Hess, S. & Daly, A. (2010), *Choice modelling: the state-of-the-art and the state-of-practice: Proceedings from the Inaugural International Choice Modelling Conference*, Emerald Group Publishing Limited.

Hess, S. & Rose, J. M. (2012), 'Can scale and coefficient heterogeneity be separated in random coefficients models?', *Transportation* **39**(6), 1225–1239.

Hess, S. & Train, K. (2017), 'Correlation and scale in mixed logit models', *Journal of Choice Modelling* **23**, 1–8.

Hole, A. R. & Kolstad, J. R. (2012), 'Mixed logit estimation of willingness to pay distributions: A comparison of models in preference and wtp space using data from a health-related choice experiment', *Empirical Economics* **42**(2), 445–469.

Hole, A. R. et al. (2006), 'Small-sample properties of tests for heteroscedasticity in the conditional logit model', *Economics Bulletin* **3**(18), 1–14.

Iversen, T. & Lurås, H. (2002), 'Waiting time as a competitive device: an example from general medical practice', *International Journal of Health Care Finance and Economics* **2**(3), 189–204.

Iversen, T. & Lurås, H. (2011), 'Patient switching in general practice', *Journal of health economics* **30**(5), 894–903.

Iversen, T. & Ma, C. A. (2011), 'Market conditions and general practitioners' referrals', *International journal of health care finance and economics* **11**(4), 245.

Iversen, T. & Ma, C. A. (2022), 'Technology adoption by primary care physicians', *Health Economics* **31**(3), 443–465.

Jack, W. (2005), 'Purchasing Health Care Services from Providers with Unknown Altruism', *Journal of Health Economics* **24**, 73–93.

Jefferson, L., Bloor, K., Birks, Y., Hewitt, C. & Bland, M. (2013), 'Effect of physicians' gender on communication and consultation length: a systematic review and meta-analysis', *Journal of Health Services Research & Policy* **18**(4), 242–248.

Jefferson, L., Bloor, K. & Hewitt, C. (2015), 'The effect of physician gender on length of patient consultations: observational findings from the uk hospital setting and synthesis with existing studies', *Journal of the Royal Society of Medicine* **108**(4), 136–141.

Jessie, D. T. & Saari, D. G. (2016), 'From the luce choice axiom to the quantal response equilibrium', *Journal of Mathematical Psychology* **75**, 3–9.

Kahneman, D. & Tversky, A. (1984), 'Choices, values, and frames.', *American Psychologist* **39**(4), 341.

Kilminster, S., Downes, J., Gough, B., Murdoch-Eaton, D. & Roberts, T. (2007), 'Women in medicine- is there a problem? a literature review of the changing gender composition, structures and occupational cultures in medicine', *Medical Education* **41**(1), 39–49.

Léger, P. T. (2008), Physician Payment Mechanisms, *in* M. Lu & E. Jonsson, eds, 'Financing Health Care: New Ideas for a Changing Society', Wiley-VCH Press, Weinheim (Germany), pp. 149–176.

Levinson, W. & Lurie, N. (2004), 'When most doctors are women: what lies ahead?', *Annals of Internal Medicine* **141**(6), 471–474.

Li, J. (2018), 'Plastic surgery or primary care? altruistic preferences and expected specialty choice of us medical students', *Journal of Health Economics* **62**, 45–59.

Li, J., Casalino, L. P., Fisman, R., Kariv, S. & Markovits, D. (2022), 'Experimental evidence of physician social preferences', *Proceedings of the National Academy of Sciences* **119**(28), e2112726119.

Li, J., Dow, W. H. & Kariv, S. (2017), 'Social preferences of future physicians', *Proceedings of the National Academy of Sciences* **114**(48), E10291–E10300.

Liu, T., Ma, C. A. & Mak, H. Y. (2018), 'Incentives for motivated experts in a partnership', *Journal of Economic Behavior & Organization* **152**, 296–313.

Louviere, J. J. & Eagle, T. (2006), Confound it! that pesky little scale constant messes up our convenient assumptions, *in* 'Sawtooth Software Conference', Sawtooth Software Inc.

Louviere, J., Street, D., Carson, R., Ainslie, A., Deshazo, J., Cameron, T., Hensher, D., Kohn, R. & Marley, T. (2002), 'Dissecting the random component of utility', *Marketing letters* **13**(3), 177–193.

Luce, R. D. (1959), *Individual Choice Behavior a Theoretical Analysis*, Oxford, England: John Wiley.

Lund, K. (2022), 'Antallet af læger kommer til at stige kraftigt de kommende år - langt mere end befolkningstallet', `https://sundhedspolitisktidsskrift.dk/nyheder/6420-antallet-af-laeger-de-kommende-ar-kommer-til-at-stige-kraftigt.html`. Accessed: 2022-12-8.

Ma, C. A. (1994), 'Health Care Payment Systems: Cost and Quality Incentives', *Journal of Economics and Management Strategy* **3**, 93–112.

McFadden, D. (1974), Conditional logit analysis of qualitative choice behavior, *in* P. E. Zarembka, ed., 'Frontiers in Econometrics', Academic Press, New York, pp. 105–142.

McFadden, D. (2001), 'Economic choices', *American Economic Review* **91**(3), 351–378.

McFadden, D., Machina, M. J. & Baron, J. (1999), Rationality for economists?, *in* F. B. & M. C.F., eds, 'Elicitation of preferences', Springer, pp. 73–110.

McKelvey, R. D. & Palfrey, T. R. (1995), 'Quantal response equilibria for normal form games', *Games and Economic Behavior* **10**(1), 6–38.

McKinstry, B. (2008), 'Are there too many female medical graduates? Yes', *BMJ* **336**(7647), 748–748.

Ministry of Health in People's Republic of China, ed. (2011), *China Health Statistic Yearbook*, Peking Union Medical College Press, Beijing.

Niederle, M. (2016), Gender, *in* J. H. Kagel & A. E. Roth, eds, 'Handbook of Experimental Economics', Vol. 2, Princeton university press, pp. 481 – 553.

OECD (2020), *OECD Health Statistics 2020*, OECD. (Accessed July 20 2022.).
URL: *https://www.oecd.org/health/health-data.htm*

Riska, E. (2001), 'Towards gender balance: but will women physicians have an impact on medicine?', *Social Science & Medicine* **52**(2), 179–187.

Roter, D. L., Geller, G., Bernhardt, B. A., Larson, S. M. & Doksum, T. (1999), 'Effects of obstetrician gender on communication and patient satisfaction', *Obstetrics & Gynecology* **93**(5), 635–641.

Roter, D. L. & Hall, J. A. (2004), 'Physician gender and patient-centered communication: a critical review of empirical research', *Annu. Rev. Public Health* **25**, 497–519.

Samuelson, P. A. (1938), 'A note on the pure theory of consumer's behaviour', *Economica* **5**(17), 61–71.

Scarpa, R., Thiene, M. & Train, K. (2008), 'Utility in willingness to pay space: a tool to address confounding random scale effects in destination choice to the alps', *American Journal of Agricultural Economics* **90**(4), 994–1010.

Scott, A. (2000), Economics of general practice, *in* A. J. Culyer & J. P. Newhouse, eds, 'Handbook of Health Economics', Vol. 1, Elsevier, pp. 1175–1200.

Sen, A. (1997), 'Maximization and the act of choice', *Econometrica* **65**, 745–779.

Shaffer, J. P. (1995), 'Multiple hypothesis testing', *Annual review of psychology* **46**(1), 561–584.

Sladek, R. M., Bond, M. J., Huynh, L. T., Chew, D. P. & Phillips, P. A. (2008), 'Thinking styles and doctors' knowledge and behaviours relating to acute coronary syndromes guidelines', *Implementation Science* **3**(1), 1–8.

Sladek, R. M., Bond, M. J. & Phillips, P. A. (2010), 'Age and gender differences in preferences for rational and experiential thinking', *Personality and Individual Differences* **49**(8), 907–911.

Ström, M. (2021), 'Andelen kvinnor ökar inom många läkarspecialiteter', `https://lakartidningen.se/aktuellt/nyheter/2021/08/andelen-kvinnor-okar-inom-manga-lakarspecialiteter/`. Accessed: 2022-12-8.

Swait, J. & Louviere, J. (1993), 'The role of the scale parameter in the estimation and comparison of multinomial logit models', *Journal of Marketing Research* **30**(3), 305–314.

Swait, J. & Marley, A. A. (2013), 'Probabilistic choice (models) as a result of balancing multiple goals', *Journal of Mathematical Psychology* **57**(1-2), 1–14.

The Norwegian Medical Association (2022), 'Om leger i norge', https://www.legeforeningen.no/om-oss/legestatistikk/om-leger-i-norge/#53301. Accessed: 2022-12-8.

Train, K. E. (2009), *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge (UK).

Train, K. E. & Weeks, M. (2005), Discrete choice models in preference space and willingness-to-pay space, *in* R. Scarpa & A. Alberini, eds, 'Applications of simulation methods in environmental and resource economics', Springer, pp. 1–16.

Tversky, A. (1972), 'Choice by elimination', *Journal of Mathematical Psychology* **9**(4), 341–367.

Varian, H. R. (1982), 'The nonparametric approach to demand analysis', *Econometrica* **50**, 945–973.

Varian, H. R. (1983), 'Non-parametric tests of consumer behaviour', *The Review of Economic Studies* **50**(1), 99–110.

Varian, H. R. (1991), *Goodness-of-fit for revealed preference tests*, Department of Economics, University of Michigan Ann Arbor.

Wallin, A., Swait, J. & Marley, A. (2018), 'Not just noise: A goal pursuit interpretation of stochastic choice.', *Decision* **5**(4), 253.

Wang, J., Iversen, T., Hennig-Schmidt, H. & Godager, G. (2020), 'Are patient-regarding preferences stable? evidence from a laboratory experiment with physicians and medical students from different countries', *European Economic Review* p. 103411.

Appendix A. Additional analyses

A1. Experimental parameters

Table A.1: Experimental parameters

	Payment	Var	0	1	2	3	4	5	6	7	8	9	10
I	FFS	R_{jA}(q)	0.00	1.70	3.40	5.10	5.80	10.50	11.00	12.10	13.50	14.90	16.60
		$R_{jB}(q)$	0.00	1.00	2.40	3.50	8.00	8.40	9.40	16.00	18.00	20.00	22.50
		$R_{jC}(q)$	0.00	1.80	3.60	5.40	7.20	9.00	10.80	12.60	14.40	16.20	18.30
		$R_{jD}(q)$	0.00	2.00	4.00	6.00	8.00	8.00	15.00	16.90	18.90	21.30	23.60
		$R_{jE}(q)$	0.00	1.00	2.00	6.00	6.70	7.60	11.00	12.30	18.00	20.50	23.00
	CAP	$R(q)$	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00
II	FFS,CAP	c(q)	0.00	0.10	0.40	0.90	1.60	2.50	3.60	4.90	6.40	8.10	10.00
III	FFS	$\pi_{jA}(q)$	0.00	1.60	3.00	4.20	4.20	8.00	7.40	7.20	7.10	6.80	6.60
		$\pi_{jB(q)}$	0.00	0.90	2.00	2.60	6.40	5.90	5.80	11.10	11.60	11.90	12.50
		$\pi_{jC(q)}$	0.00	1.70	3.20	4.50	5.60	6.50	7.20	7.70	8.00	8.10	8.30
		$\pi_{jD(q)}$	0.00	1.90	3.60	5.10	6.40	5.50	11.40	12.00	12.50	13.20	13.60
		$\pi_{jE(q)}$	0.00	0.90	1.60	5.10	5.10	5.10	7.40	7.40	11.60	12.40	13.00
	CAP	$\pi(q)$	12.00	11.90	11.60	11.10	10.40	9.50	8.40	7.10	5.60	3.90	2.00
IV	FFS,CAP	B1k(q)	0.00	0.75	1.50	2.00	7.00	10.00	9.50	9.00	8.50	8.00	7.50
		$B_{2k}(q)$	0.00	1.00	1.50	10.00	9.50	9.00	8.50	8.00	7.50	7.00	6.50
		$B_{3k}(q)$	0.00	0.75	2.20	4.05	6.00	7.75	9.00	9.45	8.80	6.75	3.00

Note: This table shows all experimental parameters. $R_{jk}(q)$ denotes physicians' payment for patient type j and illness k. Under FFS, $R_{jk}(q)$ varies with illnesses k and increases in q, whereas under CAP, $R_{jk}(q)$ remains constant. The costs for providing medical services $c_{jk}(q)$ increase in q and are the same under all experimental conditions. The physicians' profit $\pi_{jk}(q)$ is equal to $R_{jk}(q) - c_{jk}(q)$. $B_{jk}(q)$ denotes the patient benefit for the three patient types $j = 1, 2, 3$ held constant across conditions.

A2. Additional empirical results and robustness checks

Table A.2: Differences between females and males in quantity choices; *p*-values from Mann-Whitney-U tests for the 30 patients. H_0: gender difference is zero.

PATIENT #	FFS	CAP
1	0.0595	0.4662
2	0.1622	0.5252
3	0.7929	0.3034
4	0.4965	0.2268
5	0.2815	0.0418
6	0.1248	0.2888
7	0.4714	0.4450
8	0.9595	0.2056
9	0.4672	0.7065
10	0.7533	0.6945
11	0.1144	0.8197
12	0.6046	0.9518
13	0.9304	0.4206
14	0.1447	0.4220
15	0.7325	0.3139

Table A.3: Analysis of quantity choices. Estimation results from a linear regression model with random effects. N=277 subjects, T=30 patients

Dependent variable: chosen q	Estimate	Robust Std. Err.†
FEMALE	-0.07	(0.07)
DOCTOR	-0.03	(0.08)
EXPERIENCE	-0.07	(0.10)
PATIENT		
FFS_2	1.56	(0.10) ***
FFS_3	1.26	(0.09) ***
FFS_4	1.55	(0.10) ***
FFS_5	1.90	(0.11) ***
FFS_6	-0.44	(0.08) ***
FFS_7	0.61	(0.13) ***
FFS_8	0.23	(0.13)
FFS_9	0.74	(0.14) ***
FFS_{10}	0.52	(0.17) ***
FFS_{11}	1.24	(0.08) ***
FFS_{12}	1.86	(0.08) ***
FFS_{13}	1.92	(0.08) ***
FFS_{14}	1.94	(0.08) ***
FFS_{15}	2.26	(0.09) ***
CAP_1	-0.39	(0.09) ***
CAP_2	-0.41	(0.10) ***
CAP_3	-0.25	(0.10) **
CAP_4	-0.14	(0.08)
CAP_5	-0.23	(0.09)
CAP_6	-1.68	(0.09) ***
CAP_7	-1.59	(0.08) ***
CAP_8	-1.56	(0.10) ***
CAP_9	-1.59	(0.10) ***
CAP_{10}	-1.58	(0.10) ***
CAP_{11}	0.58	(0.11) ***
CAP_{12}	0.55	(0.11) ***
CAP_{13}	0.61	(0.11) ***
CAP_{14}	0.66	(0.11) ***
CAP_{11}	0.70	(0.10) ***
constant	5.06	(0.09) ***
rho = 0.13	(fraction of variance due to u_i)	

† Clustered at the level of the individual decision-maker.***(**)[*] indicate statistically significant parameters, with p-value <0.0001(<0.001)[<0.01] in a two-sided test.

Table A.4: Analysis of quantity choices. Results from maximum likelihood estimation of an ordinal regression model with random effects. N=277 subjects, T=30 patients.

Dependent variable: chosen q	FFS		CAP	
	Estimate	Robust Std. Err.†	Estimate	Robust Std. Err.†
FEMALE	-0.16	0.15	0.02	0.12
DOCTOR	-0.10	0.15	0.03	0.12
EXPERIENCE	-0.09	0.07	-0.10	0.07
PATIENT #				
2	7.65	0.15 ***	6.91	0.15 ***
3	7.19	0.15 ***	7.12	0.15 ***
4	7.51	0.15 ***	7.21	0.15 ***
5	8.06	0.16 ***	7.13	0.15 ***
6	5.17	0.13 ***	5.04	0.14 ***
7	6.45	0.15 ***	5.11	0.14 ***
8	5.99	0.14 ***	5.09	0.14 ***
9	6.56	0.15 ***	5.03	0.14 ***
10	6.48	0.16 ***	5.04	0.14 ***
11	7.26	0.16 ***	8.64	0.16 ***
12	8.09	0.15 ***	8.59	0.16 ***
13	8.18	0.15 ***	8.71	0.16 ***
14	8.16	0.15 ***	8.74	0.16 ***
15	8.77	0.16 ***	8.83	0.16 ***
Cutoffs				
cut_1	3.00	0.14 ***	2.93	0.12 ***
cut_2	3.08	0.14 ***	3.27	0.13 ***
cut_3	3.18	0.15 ***	3.51	0.13 ***
cut_4	4.18	0.15 ***	5.68	0.14 ***
cut_5	4.62	0.16 ***	6.40	0.15 ***
cut_6	6.13	0.16 ***	8.41	0.16 ***
cut_7	7.12	0.17 ***	9.90	0.17 ***
cut_8	9.03	0.18 ***	11.88	0.20 ***
cut_9	10.23	0.19 ***	12.46	0.21 ***
cut_{10}	10.89	0.19 ***	13.03	0.24 ***
var($u_{i\text{FFS}}$)	= 1.22			
var($u_{i\text{CAP}}$)	= 0.78			
cov($u_{i\text{FFS}}$, $u_{i\text{CAP}}$)	= -0.43			
Log likelihood	=-16254.871			

† Clustered at the level of the individual decision-maker.

***(**)[*] indicate statistically significant parameters with p-value <0.0001(<0.001)[<0.01] in a two-sided test.

The model is estimated by means of the gsem module in STATA 16.

A3. Interpreting the effects on the degree of randomness in Table 3.

While the degree of randomness in treatment choices does not differ significantly between genders, it might be interesting to give an interpretation of the absolute values of the point estimates of θ_{DOCTOR}, $\theta_{\text{EXPERIENCE}}$ and θ_{FEMALE}. The interpretation is that the gender difference is smaller than the difference caused by having experienced 15 additional decisions in the laboratory. The standard error of θ_{FEMALE} is smaller, and the confidence interval narrower, than the corresponding estimates for θ_{DOCTOR} and $\theta_{\text{EXPERIENCE}}$. A normalization frequently applied for providing a relative scale when comparing the degree of randomness across groups or choice occasion is achieved by reformulating Equation (6) as

$$\tilde{F}_{njt} = \Omega_{nt}[\alpha_n ln(B_{jt}) + (1-\alpha_n)ln(\pi_{jt})] + (1-\Omega_{nt})[a_j + \varepsilon_{njt}] \quad \alpha_n \in (0,1)\ \forall n \ , \tag{9}$$

where Ω_{nt} is defined by:

$$\Omega_{nt} = \frac{1}{1+\sigma_{nt}} \ , \tag{10}$$

and σ_{nt} is the scale parameter defined in Equation (7). The elements in the first bracket of (9)are the rational part of the individuals' objective, which is the individual's deterministic utility as a function of health benefits B_{jt} and profit π_{jt}, and $\Omega_{nt} \in (0,1)$ denotes the weight assigned to this part of the objective. Hence, with the normalization in (9) Ω_{nt} represents *degree of determinism* while $1-\Omega_{nt}$ represents the *degree of randomness* in behavior.

In order to illustrate how a gender difference in θ would have translated to differences in Ω_{nt}, we compute the Ω_{nt}-estimates for doctors and medical students in the two parts of the experiment, $t=1$ or $t=2$ (EXPERIENCE=0 or 1). We present the results in Table A.5.

Table A.5: Variation in the *degree of determinism* over patients and subject group. Ω_{nt} estimates are obtained by using Equation (10) and the maximum likelihood estimates in Table 3

	Doctors		Students	
Patient	**t=1** (EXPERIENCE= 0)	**t=2** (EXPERIENCE=1)	**t=1** (EXPERIENCE=0)	**t=2** (EXPERIENCE=1)
FFS_1	0.63	0.73	0.75	0.83
FFS_2	0.59	0.70	0.72	0.81
FFS_3	0.64	0.74	0.76	0.84
FFS_4	0.64	0.75	0.76	0.84
FFS_5	0.57	0.68	0.70	0.79
FFS_6	0.80	0.87	0.88	0.92
FFS_7	0.54	0.66	0.68	0.78
FFS_8	0.62	0.73	0.75	0.83
FFS_9	0.57	0.69	0.70	0.80
FFS_{10}	0.46	0.58	0.60	0.71
FFS_{11}	0.67	0.77	0.78	0.85
FFS_{12}	0.62	0.72	0.74	0.82
FFS_{13}	0.75	0.83	0.84	0.90
FFS_{14}	0.66	0.76	0.78	0.85
FFS_{15}	0.75	0.83	0.84	0.90
$CAP_{1\text{-}5}$	0.56	0.67	0.69	0.78
$CAP_{6\text{-}10}$	0.58	0.69	0.71	0.80
$CAP_{11\text{-}15}$	0.72	0.81	0.82	0.88

Note: t=1 (2) when subjects decide in Part 1 (2) of the experimental session.

When comparing the differences in Ω between students and doctors, and over the decision-maker's level of experience, we keep in mind that the interpretation of the point estimates of the three θ parameters is that the gender difference in Ω is smaller than the change in Ω caused by having more experience. We remember that Ω represents the *degree of determinism*, reflecting the degree in which choices are driven by utility differences, while $1 - \Omega$ measures the degree in which choices are driven by factors that are irrelevant to utility. We observe in Table A.5 that for decisions made in the second half of the experiment, treatment choices are more influenced by utility differences, and less influenced by irrelevant aspects: In the first line of Table A.5, we observe that for doctors deciding for patient 1 in FFS, Ω equals 0.63 if the decision maker is *less* experienced ($t = 1$, decision occurs in sessions where FFS precedes CAP). Ω rises to 0.73 when the decision maker is *more* experienced ($t = 2$, decision occurs in sessions where CAP precedes FFS). The same result also applies to decisions made by students, and we observe that when students decide for patient 1 in FFS, Ω rises from 0.75 for the less experienced student ($t = 1$) to 0.83 for the more experienced student ($t = 2$).[15] The variation in Ω over patients is substantial, ranging from the lowest estimate of 0.46 for patient 10 in FFS, to the highest estimates, 0.90, which are found for patients 13 and 15 in FFS.

Appendix B: Experimental design

B1: Decision situation

Each participant in our experiment acts in the role of a physician and is assumed to be concerned about her own profit π as well as about the patient benefit B. He/she has to choose a quantity of medical services for a given patient whose health benefit is determined by that choice. Each physician i decides on the quantity of medical services $q \in 0, 1, ..., 10$ for three patient types ($j = 1, 2, 3$) with five abstract illnesses ($k = A, B, C, D, E$). The combination of patient type and illness characterizes a specific patient $1A, 1B, 1C, ..., 3D, 3E$. Patient types differ in the health benefit they gain from the medical services ($B_{1k}(q)$, $B_{2k}(q)$, $B_{3k}(q)$). A common characteristic of $B_{jk}(q)$ is a global optimum q^*_{jk} on the quantity interval [0,10] that yields the highest benefit to patients of type j for illnesses k. The level of health benefit patients receive from optimal treatment is nearly the same for all three patient types, only the quantity of medical services differs to get there. The three types of patients reflect the patients' different states of health (intermediate, good, bad).

To illustrate the physicians' task, Figure 1a provides the decision screens for patient 1C under FFS, whereas Figure 1b shows the decision screen for the same patient under CAP. Columns 1 to 6 of the screen, respectively, indicate: (1-2) medical services and the corresponding quantities; (3) physician's remuneration, increasing in the quantity of medical services under FFS (Figure 1a), whereas under CAP the remuneration corresponds to a lump-sum payment per patient (Figure 1b); (4) costs of medical services that are constant across patient types in both parts of the experiment; (5) physician's profit (remuneration minus costs); (6) patient benefit.

B2: Parameters

Under FFS, physicians' remuneration increases in q, and remuneration differs with illnesses, $R_{jA}(q), R_{jB}(q), \ldots, R_{jE}(q)$. Physicians are paid a lump sum of 12 Token per patient under CAP, which was set close to the mean of the maximum profits a subject could achieve under FFS when averaging over patients. For an overview of all payment parameters, see panel I in Table A1 in Appendix A.

The patient benefit $B_{jk}(q)$ varies across patient types. The quantities that maximize patient benefit are $q^*_{1k} = 5$, $q^*_{2k} = 3$ and $q^*_{3k} = 7$ for patient types 1, 2, and 3, respectively with the highest level of health benefit from optimal treatment being nearly the same for all three patient types. Patient benefit $B_{jk}(q)$ is shown in panel IV of Table A1.

Further parameters relevant for physicians' decisions are costs $c_{jk}(q)$ and, particularly, profit $\pi_{jk}(q)$; see panels II and III of Table A.1. Under both payment systems, physicians have to bear costs

[15] An interesting finding which Wang et al. (2020) did not report in their paper, is that Ω_{nt} also rises as decision makers acquire experience with the current payment scheme. This can most easily be seen in CAP where Ω rises from one patient (choice occasion) to the next without exceptions.

$c_{jk}(q) = 1/10 \times q^2$. Under CAP, profits are the same for all illnesses, and the profit-maximizing quantity, $\hat{q}$, is 0 for all patients, jk. Under FFS, profits vary across illnesses because remuneration differs while costs are kept constant. The profit-maximizing quantity, $\hat{q}$, is 10 for all patients, jk, except for those with illness A, (i.e., patients 1A, 2A and 3A) as $\hat{q}_{jA} = 5$. For patient 1A, $\hat{q} = q^* = 5$. For the sake of simplicity, the patients are numbered from 1 to 15.

Figure 1a: Decision screen for patient 1C under FFS

Patient type 1/Illness C

Medical services	Quantity	Your Remuneration (in Taler)	Your Cost (in Taler)	Your Profit (in Taler)	Patient benefit (in Taler)
none	0	0.00	0.00	0.00	0.00
Service C1	1	1.80	0.10	1.70	0.75
Service C1, Service C2	2	3.60	0.40	3.20	1.50
Service C1, Service C2, Service C3	3	5.40	0.90	4.50	2.00
Service C1, Service C2, Service C3, Service C4	4	7.20	1.60	5.60	7.00
Service C1, Service C2, Service C3, Service C4, Service C5	5	9.00	2.50	6.50	10.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6	6	10.80	3.60	7.20	9.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7	7	12.60	4.90	7.70	9.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8	8	14.40	6.40	8.00	8.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9	9	16.20	8.10	8.10	8.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9, Service C10	10	18.30	10.0	8.30	7.50

Your Decision

Please indicate the quantity of medical services you want to provide

OK

Figure 1b: Decision screen for patient 1C under CAP

Patient type 1/Illness C

Medical services	Quantity	Your Remuneration (in Taler)	Your Cost (in Taler)	Your Profit (in Taler)	Patient benefit (in Taler)
none	0	12.00	0.00	12.00	0.00
Service C1	1	12.00	0.10	11.90	0.75
Service C1, Service C2	2	12.00	0.40	11.60	1.50
Service C1, Service C2, Service C3	3	12.00	0.90	11.10	2.00
Service C1, Service C2, Service C3, Service C4	4	12.00	1.60	10.40	7.00
Service C1, Service C2, Service C3, Service C4, Service C5	5	12.00	2.50	9.50	10.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6	6	12.00	3.60	8.40	9.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7	7	12.00	4.90	7.10	9.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8	8	12.00	6.40	5.60	8.50
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9	9	12.00	8.10	3.90	8.00
Service C1, Service C2, Service C3, Service C4, Service C5 Service C6, Service C7, Service C8, Service C9, Service C10	10	12.00	10.0	2.00	7.50

Your Decision

Please indicate the quantity of medical services you want to provide

OK

Appendix C: Experiment material

C1: Instructions of the experiment

[Numbers/text in brackets refer to the conditions where doctors participate.]

{Sentences/decision screens in braces are inserted into the instructions either in condition FFS or in condition CAP.}

[[Text in double brackets refer to explanatory notes.]]

Instructions Part 1
General Information

In the following experiment, you will make a couple of decisions. Following the instructions and depending on your decisions, you can earn money. It is therefore very important that you read the instructions carefully.

You take your decisions anonymously on your computer screen. During the experiment, you are not allowed to talk to any other participant. Whenever you have a question, please raise your hand. The experimenter will answer your question in private in your cubicle. If you disregard these rules, you can be excluded from the experiment without receiving any payment. All amounts of money in the experiment are stated in Token. At the end of the experiment, your earnings will be converted into RMB at an exchange rate of 10 Token = 1 [6] RMB and paid to you in cash.

The experiment consists of two parts. We we will inform you now on the decision situation in Part 1. We will provide you with the instructions for Part 2 as soon as Part 1 has ended. Please note that your decisions in Part 1 have no influence on your decisions in Part 2 and vice versa.

Your decisions in Part 1 of the experiment
During the experiment, you are in the role of a physician. You have to make 15 decisions regarding the treatment of patients. All participants of this experiment take their decisions in the role of physicians. You decide on the quantity of medical services you want to provide for given clinical symptoms of a patient.

You decide on your computer screen where five different kinds of clinical symptoms – A, B, C, D, and E – of three different patient types – 1, 2, and 3 – will be shown one after another. For each patient you can provide 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Your remuneration is as follows:

{Condition CAP: For each patient you receive a lump-sum payment that is independent of the quantity of medical services.}

{Condition FFS: A different payment is assigned to each quantity of medical services. The payment increases in the quantity of medical services.}

While deciding on the quantity of medical services, in addition to your payment you determine the costs you incur when providing these services. Costs increase with increasing quantity provided. Your profit in Token is calculated by subtracting your costs from your payment.

A certain benefit for the patient is assigned to each quantity of medical services, the patient benefit that the patient gains from your provision of services (treatment). Therefore, your decision on the quantity of medical services not only determines your own profit, but also the patient benefit. An example for a decision situation is given on the following screen.

{Decision screen for patient 1C under FFS and CAP}

[[NOTE: The same screens as in Figures 1a and 1b in Appendix B1.]]

You decide on the quantity of medical services on your computer screen by typing an integer between 0 and 10 into the box labeled ”Your Decision”.

After all participants have taken their decisions for the respective patient you will proceed to the next patient. There are no real, but abstract patients participating in this experiment. Yet, the patient benefit, which an abstract patient receives by your providing medical services, will be beneficial for a real patient. The total amount of patient benefit determined by your 15 decisions will be provided to a patient with cancer treated in Shandong Qilu Hospital [Shandong University Cancer Hospital]. The money will be directly transferred to the patient’s in-hospital account to finance part of his/her treatment fee.

Each time you make a decision on the quantity of medical services you will be informed on your profit and the patient benefit. After you have made your 15 decisions in Part 1 of the experiment you will get to know your total profit and the corresponding total patient benefit.

Earnings in Part 1 of the experiment

After you have made your decisions in Part 1 of the experiment, your overall earnings will be calculated by summing up your profits from providing medical services to the 15 patients. This amount will be converted from Token into RMB. Your earnings of Part 1 of the experiment together with the earnings of Part 2 will be paid to you in cash at the end of the experiment (rounded to 1 Yuan).

The patient benefit gained by all 15 patients will be converted into RMB at the end of the experiment, too, and will be transferred to the real patient’s in-hospital account. To this end the experimenter and a monitor will go together to Shandong Qilu Hospital [Shandong University Cancer Hospital]. After the transfer, the signed receipt will be scanned into electronic form and will be sent to all the participants via e-mail in order to ensure the authenticity of the above process. Personal information will be blinded black to respect the patient’s privacy.

After the end of Part 2 of the experiment, one participant is randomly assigned the role of the monitor. The monitor receives a payment of 50 [200] RMB in addition to the payment from the experiment. In the end, the monitor signs a form to verify that the procedure described above was actually carried out. This form will be sent to all participants together with the receipt via e-mail.

Next, please answer some questions familiarizing you with the decision situation. After your 15 decisions, please answer some further questions on your screen.

Instructions Part 2

The experiment will now be repeated including one change. Like in Part 1 you will make 15 decisions. After these 15 decisions the experiment will end.

The General Information from Part 1 also applies for Part 2 of the experiment.

Your decisions in Part 2 of the experiment

Also in Part 2 of the experiment, you are in the role of a physician and you have to make 15 decisions regarding the treatment of patients. All participants take their decisions in the role of physicians.

You decide on the quantity of medical services you want to provide for given clinical symptoms of a patient.

Like in Part 1 you decide on your computer screen where five different kinds of clinical symptoms A, B, C, D, and E of three different patient types (1, 2, and 3) will be shown one after another. For each patient you can provide 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 medical services.

Your remuneration is as follows:
{Condition CAP: For each patient you receive a lump-sum payment that is independent of the quantity of medical services.}
{Condition FFS: A different payment is assigned to each quantity of medical services. The payment increases in the quantity of medical services.}

As in Part 1, while deciding on the quantity of medical services, in addition to your payment you determine the costs you incur when providing these services. Costs increase with increasing quantity provided. Your profit in Token is calculated by subtracting your costs from your payment.

A certain benefit for the patient is assigned to each quantity of medical services, the patient benefit that the patient gains from your provision of services (treatment). Therefore, your decision on the quantity of medical services not only determines your own profit, but also the patient benefit. An example for a decision situation is given on the following screen.

{Decision screen for patient 1C under FFS and CAP}

[[NOTE: The same screens as in Part 1. Yet, participants who saw the screen for FFS in Part 1, now see the screen for CAP – and vice versa]]

You decide on the quantity of medical services on your computer screen by typing an integer between 0 and 10 into the box labeled "Your Decision".

After all participants have taken their decisions for the respective patient you will proceed to the next patient.

Also in this part of the experiment there are no real, but abstract patients participating in this experiment. Yet, the patient benefit, which an abstract patient receives by your providing medical services, will be beneficial for a real patient. Also in the second part of the experiment the total amount of patient benefit determined by your 15 decisions will be provided to a patient with cancer treated in Shandong Qilu Hospital [Shandong University Cancer Hospital]. The money will be directly transferred to the patient's in-hospital account to finance part of his/her treatment fee.

Each time you made a decision on the quantity of medical services you will be informed on your profit and the patient benefit. After you have made your 15 decisions in Part 2 of the experiment you will get to know your total profit and the corresponding total patient benefit.

Earnings in Part 2 of the experiment
After you have made your decisions in Part 2 of the experiment, your overall earnings will be calculated by summing up your profits from providing medical services to the 15 patients. This amount will be converted from Token into RMB at the end of the experiment and will be paid to you in cash together with the earnings of Part 1 of the experiment (rounded to 1 Yuan).

The patient benefit gained by all 15 patients will be converted into RMB at the end of the experiment, too, and will be transferred to the real patient's in-hospital account. To this end the experimenter and a monitor will go together to Shandong Qilu Hospital [Shandong University University Hospital]. After the transfer, the signed receipt will be scanned into electronic form and will be sent to all the participants via e-mail in order to ensure the authenticity of the above process. Personal information will be blinded black to respect the patient's privacy. Information about the procedure has been given in Part 1 of the experiment.

Next, please answer some questions in this part of the experiment that will familiarize you with the present decision situation. After your 15 decisions, please answer some further questions on your screen.

C2: Comprehension questions prior to the experiment

Please read the instructions carefully. If you have a question, please raise your hand. The experimenter will come to you and answer your question. Have you understood the instructions?

To familiarize you with the decision situation we first ask you to answer 3 questions. We will inform you when the actual experiment starts.

Assume a physician wants to provide the quantity of 0 [10, 4] medical services for the patient above.
1 [2, 3] a) What is the remuneration?
1 [2, 3] b) What are the costs?
1 [2, 3] c) What is the profit?
1 [2, 3] d) What is the patient benefit?

The test questions are now completed. When you click on the button the experiment will start.