Health Care Manag Sci (2009) 12:351–362
DOI 10.1007/s10729-008-9089-0

Incentive compatibility in kidney exchange problems

Silvia Villa · Fioravante Patrone

Received: 24 December 2007 / Accepted: 14 October 2008 / Published online: 12 November 2008

Abstract The problem of kidney exchanges shares common features with the classical problem of exchange of indivisible goods studied in the mechanism design literature, while presenting additional constraints on the size of feasible exchanges. The solution of a kidney exchange problem can be summarized in a mapping from the relevant underlying characteristics of the players (patients and their donors) to the set of matchings. The goal is to select only matchings maximizing a chosen welfare function. Since the final outcome heavily depends on the private information in possess of the players, a basic requirement in order to reach efficiency is the truthful revelation of this information. We show that for the kidney exchange problem, a class of (in principle) efficient mechanisms does not enjoy the incentive compatibility property and therefore is subject to possible manipulations made by the players in order to profit of the misrepresentation of their private information.

Keywords Game theory · Mechanism design · Incentive compatibility · Efficient mechanisms · Incomplete information · Maximum weight matchings · Kidney transplantation · Kidney exchange

S. Villa (✉)
Dipartimento di Matematica, Università degli Studi
di Genova, Via Dodecaneso, 35, 16146 Genova, Italy
e-mail: villa@dima.unige.it

F. Patrone
Dipartimento di Ingegneria della Produzione,
Termoenergetica e Modelli Matematici, Università
di Genova, P.le Kennedy - Pad D, 16129 Genova, Italy
e-mail: patrone@diptem.unige.it

1 Introduction

Kidney transplantation is nowadays an elective treatment for many renal failures; for this reason the waiting list for kidneys from deceased donors is very long. Since a healthy donor can donate one kidney to an intended recipient, the transplantation from a living person is a viable alternative to the waiting list for patients who find a willing donor.

However it can happen that for medical reasons the willing donor is not compatible with the intended recipient (see [19]). To overcome this problem, kidney exchanges between two pairs of incompatible donor-recipient pairs such that the patient in each pair can profitably receive a kidney from the donor in the other one, have been carried out in several countries. Also exchanges involving more than two pairs have been performed.

Ethical and legal issues lie behind these acts. We shall not address these questions, even if they can be subjected to a formal scrutiny, and even if they are relevant in defining the constraints under which one has to operate (see [4, 14, 19]).

In this paper we will restrict our attention to exchanges involving only two incompatible pairs (for a detailed analysis of the case in which no restrictions on the cycles' length are imposed see [16]). Note that the restriction to incompatible pairs is suggested both on the basis of ethical issues and on the consideration of the risks connected with actions that would allow to speak truly of an organ market; on the other hand the restriction to direct exchanges is due to a practical motivation. In fact, since in most countries the consent to the donation can be withdrawn in any moment, all the surgeries involved in a fixed exchange must be carried out simultaneously and this constraint requires a considerable number of medical teams to be available at the same moment.

In the real situation, in the countries where such exchanges are performed (see e. g. [2, 3, 5, 6, 13]), usually a standard program in order to organize the exchanges is adopted. We are going to present a model of the organization of a kidney exchange program, on the basis of [3], which is however similar to the others cited above. The program of kidney exchanges is organized in several steps, in chronological order, as follows. Each transplant center has a fixed period of time to collect a group of incompatible patient-donor pairs. More precisely it is probable that each patient presents more than one willing donor, so that the transplant centers collect pairs of the type patient-set of donors. After the enrollment period, a centralized (national) database of incompatible pairs patient-set of donors is formed. At this point a central committee decides the exchanges to be carried out using a fixed chosen mechanism. At the final stage, the list of chosen exchanges is announced and the transplants are carried out, if no withdrawals occur.

It is clear that this kind of procedure has some tracts typical of a market, being quite close to some kind of barter. So, even if there is not a complete coincidence from the point of view of interpretation (these exchanges are often referred to as an exchange of gifts), it is natural to see whether and how the literature on the exchange of indivisible goods can be applied in this setting.

In this paper we analyze some essential aspects arising in the design of the rules of the "market," using the point of view of mechanism design and game theory ([7–9, 15, 18]).

We focus on a particular mechanism, very similar to some mechanisms adopted in practice (see [3, 6]), that associates to each kidney exchange problem a collection of exchanges satisfying some natural requirements arising in this medical setting (see Section 3 for details). In particular we address the natural requirement of efficiency by choosing a maximum weight matching (see Section 3). An appropriate choice of the weights of the various exchanges should in fact favor by construction the best transplants in terms of "medical quality," at least in principle. In fact, one of the problems is that efficiency can be impaired by a strategic behavior of the patients, who may have incentives to use their own private information in their favor.

Using the terminology of mechanism design, a mechanism in which the patients are induced to truthfully reveal their private information is called incentive compatible (see Section 2 for a brief introduction on mechanism design). We will show with three counterexamples that the patients can behave strategically in order to bend the maximum weight mechanism in their favor, so that the chosen mechanism is not incentive compatible. This makes our situation completely different from the case studied in [17], in which it is shown that declaring the truth is a dominant strategy.

Let us now clarify what kind of private information each player possesses. The set of willing donors is certainly part of it. Moreover, we will assume that each patient, (and/or each patient's doctor), knows who are the donors of the other pairs involved in the system who are compatible with him, and also that this information is not available to the other patients. We will see in Section 6 that both these aspects are relevant in determining the strategies used by the patients. In fact, the possibility of strategic manipulation arises at this point, because without involving any illicit behavior, a patient can decide to hide some of the donors.

In the next sections we will examine this possibility, studying the property of incentive compatibility at various levels of information. We will consider three different scenarios, corresponding to three different levels of information, namely complete information, partially incomplete information, and incomplete information. Even if we study them independently, as three one-shot games, and no dynamic analysis is carried out in this paper, doing it would be interesting, since the motivation for the scenarios (or level of information) that we consider comes from steps that are found in some of the existing organizations which carry out such exchange program, so that they could be seen as stages of the intrinsically dynamic real situation.

Going more into the details, the case of incomplete information corresponds to a situation in which each player does not know anything of the others. He does not even know how many people are taking part in the exchange. He has anyway a significant amount of information coming from public sources concerning the medical (or genetic) characteristics of the other participants: he knows, or may know (or his doctor knows), the probability distribution over the relevant parameters for compatibility (e.g. blood and tissue types, the age, and the weight). He knows what is the probability of a positive cross-match, a medical condition forbidding the transplant, between two persons. With this information, he is able to construct what we will call the "compatibility matrix" between types. In this work we make two basic assumptions: the first one is that the quality of an exchange depends only on the types of the two individuals involved and the second one is that the "rules" adopted for computing the compatibility between a donor and a recipient are common knowledge.

The first assumption agrees with the choice to measure the quality of an exchange taking into account only the quality, identified with the compatibility level, of the two transplants constituting the exchange. Certainly this is not the only possible or reasonable choice: a natural modification of our method gives rise to models in which factors other than the transplant's quality are taken into account, like the gravity of the clinical situation of the involved patients or the time spent on the waiting list. The second assumption is almost unavoidable. In fact the mechanism should be common knowledge, since a fundamental property required on a mechanism is its accountability, and in the medical setting this property assumes even a deeper meaning than usual. So it would not be acceptable that the criteria used by the transplant center in order to decide which exchanges are preferred to the others are not common knowledge. Note that our assumption simply implies that given all the relevant parameters everyone knows how to determine the compatibility between a donor and a recipient, and does not imply that the values of the relevant parameters in each situation are common knowledge.

The case of partial incomplete information can be seen as a model of the situation in which each doctor knows each set of donors of the involved patients. We remark that this occurs in practice, since the evaluation of compatibility, being not an easy task, is likely to be left to the doctor of the potential recipient. Moreover we implicitly assume that the information sets of the patients coincide with the information sets of their doctors. We simply note that this amount of information can be used by the players in order to get better results for themselves. Since the consent to the donation can be retired in any moment, after this release of information maybe some patient prefers to declare as unavailable some donor previously declared as available.

In the case of complete information, every patient knows exactly the relevant parameters corresponding to the other patients and donors involved in the exchange program.

We will show that at each information level there are problems and cases in which it is convenient for a patient not to declare all the potential donors.

Our model is different from the one in [17]: the difference with respect to [17] lies mainly in the fact that we put some weights that make some exchanges better than others from the medical perspective, while this was not present in the model [17]. Our hypothesis agrees with the point of view of the european surgeons that suggest a link between donor and patient characteristics and the resulting quality of the transplant ([11, 12]) and is in contrast with the assumption of american surgeons that assume that the patient is indifferent among compatible healthy kidneys. The approach of [16] is more similar to our model, but it essentially differs in the proposed solution, because no restriction on the cycles length is imposed.

In the presentation of our analysis, after some preliminaries on mechanism design (see Section 2), we start from the simpler case to model, in which there is complete information (see Sections 3 and 4), and in Sections 5 and 6 we consider the cases of incomplete information and partial incomplete information.

2 Some preliminaries on mechanism design

The typical situation studied in the mechanism design theory involves a group of agents, a designer, and a set of possible outcomes. The goal of the designer is the realization of an outcome prescribed by a given social choice function, which is defined starting from the preferences of the agents on the set of possible outcomes. Usually the problem is that the designer is at an informational disadvantage with respect to the agents, in the sense that he has a partial information about their preferences. A central problem in the theory of mechanism design is whether the designer has the possibility to invent an appropriate mechanism. Appropriateness means that the interaction between the agents, given their real preferences, produces an outcome realizing the social choice function. Since in the standard situation more than one agent is involved, to choose a mechanism means to choose a game form that the agents have to play. The main suggestion coming from the theory of implementation is that incentives should be given to the agents in order to obtain the desired outcome as a consequence of strategic behavior, in such a way that the desired outcome coincides with the outcome obtained in correspondence of an appropriately defined solution concept for the chosen game.

Corresponding to different solution concepts possibility and impossibility results have been obtained. The most demanding way of implementation is the one in dominant strategies: in this case the selected outcome has to be obtained when the agents choose dominant strategies in the game they are playing. The most famous result of implementation in dominant strategies is the Gibbard-Satterthwaite impossibility theorem that states that if there are at least three possible outcomes and the designer does not know the private information of the players the only social choice functions that can be implemented in dominant strategies are dictatorial.

Weakening the requirements on the solution, for instance passing to Nash equilibrium, it is possible to get positive results. In our paper, since we are in the case of a static game of incomplete information also between the agents, the underlying implicit solution concept is the Bayesian Nash equilibrium (for the precise definitions see [10] and the Appendix).

A key result in the theory of mechanism design is the revelation principle (see [1, 10, 18] and the Appendix) for general bayesian games, stating that it is possible to restrict the attention to the so called direct mechanisms, which simply ask each agent to report his private information and choose an outcome based on this.

In this setting one of the necessary conditions for implementability is called incentive compatibility, which says that a social choice function must be such that truthful revelation of the private information is an equilibrium strategy.

3 Model of the kidney exchange problem in the complete information setting

Let $N = \{1, \dots, n\}$ be a group of patients, and suppose that each patient i has a set of potential donors D_i (incompatible with him). Let us denote by $\mathcal{D} = \prod_{i=1}^{n} D_i$. Throughout the paper we identify the pair (i, D_i) with the player i. Sometimes we also call patient i the player i since we will assume that the utility function of player i coincides with the utility function of the patient. We will specify later the assumptions on the functional form of the utility function. Each individual has several relevant characteristics influencing the compatibility with other people, among them we recall the blood type, the tissue type, the age, and the weight. We assume that these characteristics give rise to a finite set of types of individuals $\{t_1, \dots, t_r\} =: T$. We model this by introducing a $r \times r$ compatibility matrix C, where the entry c_{ij} expresses the compatibility level between a donor of type t_i and a patient of type t_j. In particular, $c_{ij} = 0$ means that a donor of type t_i cannot donate to a recipient of type t_j. In the real world, in addition to the compatibility between types also the so called negative cross-match is required in order to make the transplant possible. Cross-match negativity is not in general predictable starting from the types, but should be verified case by case (even if some hints in order to estimate its probability come from the similarity between the tissue type of the donor and the recipient, and from the history of the recipient). Anyway, using the compatibility matrix C and the information on the cross-match we can construct a weight matrix $W(N, \mathcal{D}) =: W$ of size $n \times n$, defined by setting

$$W_{ij} = \begin{cases} 0 & \text{if each donor in } D_i \text{ is not compatible} \\ & \text{with patient } j \\ c_{\sigma(i,j),j} & \text{otherwise,} \end{cases}$$

where $\sigma(i, j) = k$, with k the type of the donor $d \in D_i$, which is the best donor for patient j among the donors of patient i. In other words W_{ij} reflects the quality of the best possible transplant between a donor in D_i and the patient j. Note that by hypothesis $W_{ii} = 0$ for every i.

Definition 1 A kidney exchange problem is a triple $(N, \mathcal{D}, W)$.

It is easy to see that a kidney exchange problem can be interpreted as an oriented weighted graph where the patients are the nodes and the ordered pair (i, j) is an arc if and only if $W_{ij} > 0$. If this is the case, the weight of the arc (i, j) is W_{ij}.

Definition 2 We say that patients i and j are mutually compatible if $W_{ij} \cdot W_{ji} > 0$.

Definition 3 A matching is a function $m : N \to N$ such that $m(i) = j$ implies $m(j) = i$ for every $i, j \in N$ and such that $m(i) = j$ with $i \neq j$ implies that i and j are mutually compatible.

Definition 4 A mechanism is a rule which assigns a matching (or more generally a lottery over matchings) to each problem.

More precisely a mechanism is a function

$$h : (N, \mathcal{D}, W) \mapsto h(N, \mathcal{D}, W) \in L(\mathcal{M}),$$

where $L(\mathcal{M})$ is the set of lotteries over matchings.

Solving a kidney exchange problem consists in finding a suitable mechanism, i.e a rule that associates a suitable matching to each kidney exchange problem. The first problem in the setting of kidney exchanges is to clarify which requirements a solution should satisfy. It is clear that any proposed solution must be not only feasible but also efficient, since kidneys are a scarce resource that cannot be wasted. The feasibility is ensured by Definition 3. Efficiency in this setting (see [10]), can be expressed in the following way: a mechanism (matching) is efficient if no other feasible mechanism (matching) can be found that might make some other individuals better off and would certainly not make other individuals worse off.

With respect to efficiency at least two aspects have to be considered: one is the number of patients that receive a transplant and the other is the quality of the performed transplants. In [17] a maximum cardinality matching is proposed as solution, satisfying the requirement of efficiency in the case of weights 0–1. Here we focus on mechanisms giving positive probabilities only to maximum weight matchings, so that it can happen that a matching guaranteeing a smaller number of transplants of better quality is preferred to another one which prescribes a bigger number of transplants, but of less quality. It is straightforward to prove that this class of mechanisms is efficient. Anyway it is quite clear that an assessment of the appropriate weights to be used is far from being an easy and non-debatable task. However, the focus of this contribution lies elsewhere.

Usually, more than one maximum weight matching exists, and several procedures of selection are certainly possible. One procedure consists in the assignment of a certain probability to each of the maximum weight matchings; another possibility is the definition of an ordering on the set of matchings that allows to select one precise matching without resorting to the use of a chance move. For instance, looking at the patients involved, one can choose the matching that guarantees a transplant to the patients that have been waiting for a longer time. In practice there is some perplexity towards the use of stochastic mechanisms, even if it is theoretically shown that they satisfy (ex-ante) equity requirements (see [17]). We do not enter into the details, since in this paper we will always take into account examples in which the maximum weight matching is unique.

We assume that the utility of patient i of receiving a kidney from the best donor in D_j is W_{ji}, which implies that the utility of patient i under the matching m is $W_{m(i),i}$. We understand that this assumption is not a light one. A patient could have his idiosyncratic way of looking at the possible consequences of a transplant and the ways to be used for their evaluation, even if it is reasonable to assume that the utility function of a patient is just coming from a best possible assessment on the quality of the transplant. And this assessment should be dictated by scientific reasons which should represent the best available estimate, e.g. the QALY (quality adjusted life years). From the point of view of the negative results obtained in this paper, the fact that we are able to show the possibility of manipulation even in this "uniform" landscape of patients endowed with the same functional form of the utility function, is an element in favor of the relevance of the results.

4 Strategic decisions in the complete information case

Each patient faces the strategic decision of declaring his set of willing donors. In the complete information setting it is easy to show that there are situations in which a patient can benefit from the decision of stating a proper subset of D_i instead of D_i.

Example 1 To simplify the calculations suppose that each individual can be only of four possible types, and suppose that the compatibility matrix has the following structure

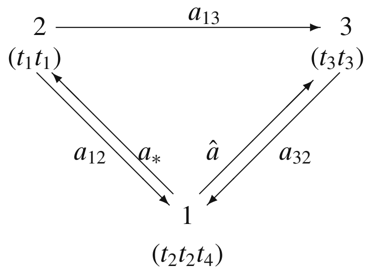
$$C = \begin{bmatrix} 0 & a_{12} & a_{13} & a_{14} \\ a_{21} & 0 & a_{23} & 0 \\ 0 & a_{32} & 0 & a_{34} \\ a_{41} & 0 & a_{43} & a_{44} \end{bmatrix},$$

with $a_{ij} > 0$, satisfying the inequalities:

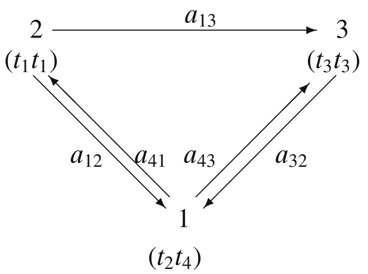
$$a_{12} > a_{32}, \quad \hat{a} + a_{32} > a_* + a_{12} \tag{1}$$

$$a_{43} + a_{32} < a_{41} + a_{12} \tag{2}$$

where $a_* = \max\{a_{21}, a_{41}\}$ and $\hat{a} = \max\{a_{43}, a_{23}\}$. Consider a kidney exchange problem with 3 patients, patient 2 is of type t_1 and has a donor of type t_1, patient 3 is of type t_3 and has a donor of type t_3, while patient 1 is of type t_2 and has two willing donors d_2 and d_4, of type t_2 and t_4 respectively. Suppose that all cross-matches are negative, so that the kidney exchange problem is represented by the following graph (the notation $(t_i t_j t_k \dots)$ means that the patient is of type t_i, the first donor is of type t_j, the second is of type t_k and so on and the number on the arc (i, j) represents the compatibility between "the best donor" in D_i and patient j):



Since patients 2 and 3 are not mutually compatible, the only feasible exchanges are $(1, 2)$ and $(1, 3)$ and we get that the maximum weight matching is $(1, 3)$ by assumption (1). Again by assumption (1) it follows that patient 1 would prefer the matching $(1, 2)$ because he prefers a donor of type t_1. Now, if patient 1 hides the donor of type t_2 the situation is the following:



and therefore by Assumption (2) the maximum weight matching is $(1, 2)$. Thus in the latter scenario patient 1 receives a transplant of quality a_{12} instead of a transplant of quality a_{32}, with $a_{12} > a_{32}$. In other words, denoting by u_1 the utility function of player 1 (depending on the strategies of all the players)

$$u_1 : 2^{D_1} \times 2^{D_2} \times 2^{D_3} :\to [0, +\infty)$$

we have

$$u_1(D_1, D_2, D_3) < u_1(\{d_2\}, D_2, D_3).$$

This is in contrast with the positive results obtained in the case of complete information in [17], where the weights are assumed to be 0 or 1. With this assumption it is proved that revealing all the set of available donors is not only an equilibrium, but also a dominant strategy both under a deterministic efficient mechanism and under a stochastic efficient mechanism (see [17] for the details).

5 The model in the incomplete information case

We now consider the incomplete information case. As in Section 3 we suppose to have a set N of patients with their donors $(D_1, \ldots, D_n)$, but now patient i does not know the compatibility matrix W, nor the set D_j for $j \neq i$.

We suppose that each patient can have at most two willing donors. Each player type is then given by a vector $t \in \mathcal{T} := T^3 \cup T^2$. This is not an essential hypothesis, but we make it in order to simplify the calculations in the following examples. The crucial assumption is the existence of a fixed common upper bound on the number of donors, and this is a very realistic assumption from the practical point of view.

We assume that a probability distribution P over possible types (i.e. on the set $\mathcal{T}$) is given and it is common knowledge. This assumption is plausible, since this information is public: the distribution of blood types is known and in the same way it is possible to obtain a probability distribution for the other characteristics. Starting from a probability p on the set T and supposing that the types of the donors and of the patients are independent, we put the product probability on the sets T^2 and T^3. If we assume that the probability of having one or two donors is given by s and $1 - s$, we put the probability P on the space $\mathcal{T}$ obtained by the combination of s, p^2 and p^3. More precisely, the probability that player k is a pair of type $(t_i, t_j) \in T^2$ is $s\, p(t_i)p(t_j)$, while the probability that he is a triple of type (t_i, t_j, t_l) is $(1 - s)p(t_i)p(t_j)p(t_l)$. The assumption on the independence of the types of the patient and the donors is quite a strong one, since it does not agree with the possibility that the patient and the donor are “blood” related, and therefore can have similar genetic characteristics. Anyway it is a good approximation of the real situation, where many donations come from an unrelated donor.

About the information of each player, it is reasonable to assume that each patient knows his type, i.e. he knows how many donors he has, and his and their relevant characteristics.

The set of actions for patient i is 2^{D_i}, therefore a strategy of patient i is a function $d_i : \mathcal{T} \to 2^{D_i}$ (see [10] and the Appendix for the definitions in the general case of games of incomplete information). We denote by $d : \tau \in \mathcal{T}^n \mapsto (d_1(\tau_1), \ldots, d_n(\tau_n))$ a profile of strategies of the patients. d univocally defines a kidney exchange problem for each type profile τ, and therefore in correspondence to τ and $d(\tau)$ a compatibility matrix is also determined. We denote this correspondence by a function

$$w : \mathcal{T}^n \times 2^{D_1} \times \cdots \times 2^{D_n} \to \mathbb{R}^{n \times n}$$

defined by setting $w(\tau, d(\tau)) = W(N(\tau)), \mathcal{D}'(\tau))$, where W is the compatibility matrix obtained in the same way as in the complete information setting with patients' types determined by τ and the set of donors $\mathcal{D}' = (d(\tau_1), \ldots, d(\tau_n))$. The main difference compared to the complete information setting is that in this case the computation of the compatibility matrix is based on the declarations of the patients. We implicitly assumed that a patient can declare a proper subset of his donors, but he can't lie on the types of the declared donors, since this false declaration would be always discovered.

In general we have seen that the compatibility matrix depends on the types, but also on the possibility of positive or negative cross-match. This information cannot be obtained from the types, since it depends on the private history of the patient, whether for instance he has already received a transplant or a blood transfusion. For this reason we assume that there is a certain probability q that the cross-match between two people is positive. In order to have an estimate of this q we refer to the real situation, in which the published probability of positive cross-match between two random people is $q = 0.11$ (see Table 1 in [19]).

Definition 5 A kidney exchange problem with incomplete information is given by a collection

$$K = \left(N, \{D_i\}_{i\in N}, \mathcal{T}^n, P, w, q\right).$$

Although function w is not independent from the data, but it is in some sense derived from them, we include it in Definition 5 since, given the data, several choices of this function can be done. Also in our setting, where we have clearly identified the target of obtaining transplants of quality as high as possible, the ideal w, which coincides with the life years guaranteed by the transplant, does exist but it is impossible to be computed. Since there are many arbitrary choices when determining an approximation of it, several distinct functions w can well represent the situation, also if they assign different weights to the arcs, and therefore they define different kidney exchange problems.

We recall that we are given a mechanism h, in this case acting on the triples $(\tau, d(\tau), w(\tau, d(\tau)))$. Clearly each profile of types and each strategy determine an element of $L(\mathcal{M})$ through the mechanism h. If h takes values in $\mathcal{M}$, it is possible to consider the function $h(\tau, d(\tau),\ w(\tau, d(\tau))) = m(\tau, d)$. The resulting matching gives the expected utility of the patients. In fact the expected utility of patient i who is of type τ_i and chooses $d_i(\tau_i) = D'_i$ is given by

$$u^e_i\left(\tau, D'_i, d_{-i}\right) = \sum_{\tau_{-i}\in\mathcal{T}_{-i}} p_i(\tau_{-i}|\tau_i)w((\tau_{-i}, \tau_i), d)_{m(\tau,d)(i),i}$$

(here the notation v_{-i} for a vector v means all the components of vector v except for the i-th).

6 Incentive compatibility under incomplete information

In this Section we show that in the case of incomplete information truthful revelation is not an equilibrium and, a fortiori, not a dominant strategy. The problem is that the probability of being matched with someone increases if the set of donors increases (see [17]), but this property does not hold for the expected utility which is not “ donor monotonic.”

We prove with two examples that the property does not hold in the incomplete information setting as depicted in Section 5, and also in an intermediate state of information between the incomplete and the complete one. We start with the incomplete information setting as modeled in Section 5 and then we modify this example in order to adapt it to the case in which the patients possess an intermediate level of information. The underlying idea is that the situation is analogous to the case of complete information: having one “good” donor is not necessarily an advantage, since this can cause as a result the reception of a kidney of lower quality.

We want to show that also in the case of incomplete information, if some types are more probable than others, it is possible to design situations in which revealing all the available donors is not a dominant strategy. In particular we show that if the strategy of all the players but player 1 is to reveal all of their donors, the best response for player 1 is not the truthful revelation of the set D_1. We do not take into account the possibility of having a positive cross-match, but this can be added without changing the essence of the example, if we suppose that the probability of having a positive cross-match is sufficiently small (see Example 3).

Example 2 As in Example 1 suppose that there are only four possible types of individuals and that the compatibility matrix is the following

$$C = \begin{bmatrix} 0 & a_{12} & 0 & a_{14} \\ a_{21} & 0 & a_{23} & 0 \\ 0 & a_{32} & 0 & a_{34} \\ a_{41} & 0 & a_{43} & a_{44} \end{bmatrix}$$

We assume that types t_1 and t_3 occur with high probability, while types t_2 and t_4 are rare. We show that there exists a kidney exchange problem in which a patient has convenience to hide a willing donor if the other players truthfully reveal the set of their donors. We make these assumptions on the compatibility matrix:

$$a_{12} > a_{32}, \quad a_{23} + a_{32} > a_* + a_{12}$$

$$a_{43} + a_{32} < a_{41} + a_{12}, \quad a_{43} + a_{12} > a_{41} + a_{32}$$

where $a_* = \max\{a_{21}, a_{41}\}$. Consider a kidney exchange problem with three patients and suppose that patient 1 is of type t_2 and has two willing donors, one of type t_2 and the other of type t_4.

Patient 1 does not know the types of the other patients and of their donors, he only knows the compatibility matrix and that the other patients with probability 1/2 have only one willing donor and with probability 1/2 have two willing donors. We observe that this choice can be easily modified, since it is not relevant in our analysis.

We assume that on the space $T = \{t_1, t_2, t_3, t_4\}$ the following probability distribution is given, for some $\epsilon \in (0, 1)$:

$$p(t_1) = p(t_3) = \frac{1-\epsilon}{2}, \qquad p(t_2) = p(t_4) = \frac{\epsilon}{2}.$$

In the first column of Table 1 we put the probability of the type profiles reported in the second column. The chosen type profiles in the second and third columns are those for which the different declaration of player 1 leads to a different outcome for patient 1 under the maximum weight matching. In the fourth column it is written the type of kidney which patient 1 receives at the maximum weight matching if he declares only donor t_4 while in the fifth there is the type of kidney that he receives declaring both his donors.

Since we supposed that $a_{12} > a_{32}$ patient 1 prefers to receive a kidney of type t_1 with respect to a kidney of type t_3.

It is clear that for each of these type profiles there is the "symmetric" one in which patients 2 and 3 exchange their roles.

Denoting by

$$T(1-\epsilon) = 2\left[\frac{(1-\epsilon)^4}{2^6} + 4\frac{(1-\epsilon)^5}{2^7} + 3\frac{(1-\epsilon)^6}{2^8}\right],$$

the expected utility of patient 1 of type t_2 declaring only donor t_4 if the other players are "honest" is

$$u_1^e(t_2, \{t_4\}) = T(1-\epsilon)a_{12} + C\left(\frac{1-\epsilon}{2}, a_{32}, a_{12}\right) + \epsilon U(\epsilon, t_2, \{t_4\}).$$

Table 1 Types' profiles leading to a different outcome for patient 1 as a consequence of different declarations

Probability	2's type	3's type	Type received from pat. 2 declaring t_4	Type received from pat. 2 declaring t_2, t_4
$\frac{(1-\epsilon)^4}{2^6}$	t_1t_1	t_3t_3	t_1	t_3
$\frac{(1-\epsilon)^5}{2^7}$	t_1t_1	$t_3t_3t_3$	t_1	t_3
$\frac{(1-\epsilon)^5}{2^7}$	t_3t_3	$t_1t_1t_1$	t_1	t_3
$\frac{(1-\epsilon)^5}{2^7}$	t_3t_3	$t_1t_1t_3$	t_1	t_3
$\frac{(1-\epsilon)^6}{2^8}$	$t_1t_1t_1$	$t_3t_3t_3$	t_1	t_3
$\frac{(1-\epsilon)^6}{2^8}$	$t_1t_1t_3$	$t_3t_3t_3$	t_1	t_3

The expected utility of patient 1 of type t_2 declaring donors t_2 and t_4 is

$$u_1^e(t_2, \{t_2, t_4\}) = T(1-\epsilon)a_{32} + C\left(\frac{1-\epsilon}{2}, a_{32}, a_{12}\right) + \epsilon V(\epsilon, t_2, \{t_2, t_4\}),$$

where C can be computed examining the most probable type profiles, and is the same in both cases, and U and V can be computed examining the remaining type profiles. If ϵ is small enough, recalling that $a_{12} > a_{32}$, it follows that

$$u_1^e(t_2, \{t_4\}) > u_1^e(t_2, \{t_2, t_4\}).$$

With the next example we show that also in the intermediate case, in which each patient (or the doctor of each patient) knows the compatibility values between the donors of the other patients and himself, but he does not know the types of the other patients, the truthful revelation of the set of donors is not an equilibrium. In this example, which is simpler to manipulate, we take into account also the role of the cross-match. This is the reason why we allow only three possible types.

Example 3 Each person can be only of 3 possible types: t_1, t_2, t_3 and each type occurs with probability 1/3 (this is not essential, but simplify the calculations). The compatibility matrix is:

$$C = \begin{bmatrix} 0 & a_{12} & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & a_{32} & 0 \end{bmatrix}.$$

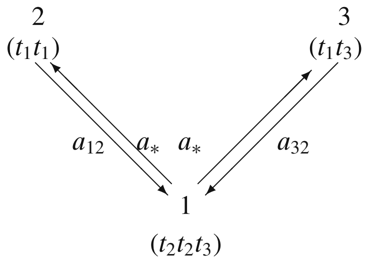
Recall that the matrix C has to be interpreted in the following sense: a donor of type t_i is incompatible with a recipient of type t_i, but $a_{ij} > 0$ does not imply that donor t_i is compatible with patient t_j. We assume that this happens only with probability $(1-q)$, with q small, while with probability q a donor t_i is incompatible with a recipient t_j, if $i \neq j$.

Consider a kidney exchange problem in which there are only 3 patients. Assume that patients 2 and 3 have only one willing donor.

Patient 1 knows that he has two willing donors. Patient 1 knows his type and the types of his donors (incompatible with him). We assume he is of type t_2 and his donors are of type t_2 and t_3 respectively. Suppose that patient 1 knows that patient 2 has a donor compatible with him, and that the compatibility is a_{12}, and also that patient 3 has a donor compatible with him and the compatibility value is a_{32}. From this knowledge he can deduce that with probability $\frac{1}{1+2q}$ patient 1 is of type t_1, with probability $\frac{q}{1+2q}$ he is of type t_2 and with probability $\frac{q}{1+2q}$ he is of type t_3. Analogously patient 1 knows that patient 2 is of type t_3 with probability $\frac{1}{1+2q}$, while with probability $\frac{q}{1+2q}$ he is of type t_1 and t_2 respectively.

The possible combinations of types of patients and donors are the ones in Table 2.

Of course, also the possibilities in which the types of 2 and 3 are interchanged have to be considered in the computation of the expected utility of patient 1. Now suppose that patient 1 declares the two donors of type t_2 and t_3 respectively. Each line of the table plus the declaration of patient 1 determines a kidney exchange problem. Let us analyze the first case for instance:



where $a_* = \max\{a_{31}, a_{21}\}$. Note that in this case the exchange between 2 and 3 is not possible, since the donor of patient 2 is certainly incompatible with recipient 3. Assuming that $a_{12} > a_{32}$ we get that with probability $1-q$ the selected matching in this case will be $(1, 2)$. If this is not possible, with probability $q(1-q)$ the selected matching will be $(1, 3)$, and if also this one is not feasible, no matchings are found in this case.

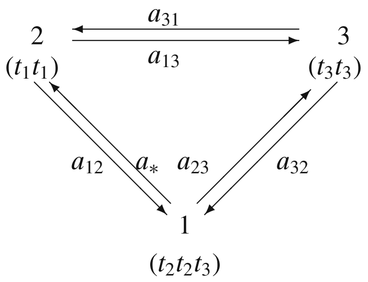
Table 2 Players' types' probabilities

Probability	Type of 2	Type of 3
$\frac{q}{(1+2q)^2}$	t_1t_1	t_1t_3
$\frac{q}{(1+2q)^2}$	t_1t_1	t_2t_3
$\frac{1}{(1+2q)^2}$	t_1t_1	t_3t_3
$\frac{q^2}{(1+2q)^2}$	t_2t_1	t_1t_3
$\frac{q^2}{(1+2q)^2}$	t_2t_1	t_2t_3
$\frac{q}{(1+2q)^2}$	t_2t_1	t_3t_3
$\frac{q^2}{(1+2q)^2}$	t_3t_1	t_1t_3
$\frac{q^2}{(1+2q)^2}$	t_3t_1	t_2t_3
$\frac{q}{(1+2q)^2}$	t_3t_1	t_3t_3

So, the expected utility of patient 1 in this case is:

$$u_1^e(\{t_2, t_3\}, t_1t_1, t_1t_3) = (1-q)a_{12} + q(1-q)a_{32}.$$

Let us see the case that occurs with the highest probability (if q is small enough): player 2 is of type t_1t_1 and player 3 is of type t_3t_3. The situation is the following:



Assuming that $a_{32} + a_{23} > a_{12} + a_* > a_{13} + a_{31}$, the selected matching will be $(1, 3)$ with probability $(1-q)$, $(1, 2)$ with probability $q(1-q)$, and finally $(2, 3)$ with probability $q^2(1-q)$.

Therefore the expected utility of 1 in this case is

$$u_1^e(\{t_2, t_3\}, t_1t_1, t_3t_3) = (1-q)(a_{32} + qa_{12}).$$

In the same manner, it is possible to compute the expected utility of patient 1 in each case. At the end, the expected utility of patient 1 declaring all his donors is an expression that looks like this:

$$u_1^e(\{t_2, t_3\}) = \frac{1}{(1+2q)^2}(1-q)a_{32} + qA(q, a_{12}, a_{32}),$$

where A is computed taking into account all the possibilities included in Table 2 and the symmetric ones. It is clear that if q is sufficiently close to 0, then the expected utility for 1 in this case is close to a_{32}.

If instead donor 1 decides to declare only the donor of type t_3 he gets:

$$u_1^e(\{t_3\}, t_1t_1, t_3t_3) = \frac{1}{(1+2q)^2}(1-q)a_{12} \\ +qB(q, a_{12}, a_{32}),$$

where again B is computed in the same way as A. Now, if q is sufficiently close to 0, the expected utility of patient 1 declaring only one of his donors is close to a_{12}. Since we assumed that $a_{12} > a_{32}$, this shows that declaring the whole set of donors is not an equilibrium strategy for patient 1.

7 Concluding remarks

Kidney exchanges between incompatible pairs of donors and patients are beginning to be performed around the world. A centralized organization of the exchanges requires a formal analysis of the chosen allocation mechanism in order to correctly assess efficiency and other desirable properties, as incentive compatibility.

We have analyzed the property of incentive compatibility in the kidney exchange program from the point of view of the mechanism design. We studied three different scenarios corresponding to three different information levels. The inspiration for considering them comes from the observation that such scenarios occur at a certain moment in some of the existing practical implementations of kidney exchange programs.

i) the first one is the incomplete information case in which the patients do not know anything but the information that can be obtained from a public source (on the distribution of blood types for instance);
ii) the second one is again a situation of incomplete information, but each patient knows the compatibility between himself and the other patients' donors;
iii) the last one is the case of complete information: each patients knows the compatibility between himself and the other patients' donors and also the compatibility of the other patients with his donors.

We found out that in each of the three scenarios the mechanism choosing a maximum weight matching does not satisfy the incentive compatibility constraints.

In other words, if efficiency is identified with the medical quality of the transplants, and the social welfare function is the sum of the utility of the patients, we proved that an efficient mechanism satisfying the incentive compatibility property does not exist (see the Appendix).

As we saw, this means that the mechanism lends itself to manipulations operated by the players, since they can use the information they possess in order to obtain a better kidney for themselves, with respect to the kidney allocated through the mechanism. Moreover, as a consequence, this behavior invalidates efficiency.

Since the strategic behavior of each patient is limited to the concealment of some donor, in the practical application of the mechanism, even if the strategic behavior at the first stage (in the incomplete information case) cannot be controlled, it is essential to try to render the manipulations as difficult as possible, by limiting for instance the release of information from the doctors to the patients, so that no donors are excluded from the program in the intermediate steps, except for the ones not declared at the first one.

Finally we remark that the examples studied here are clearly "fictitious," even if they preserve the salient features of the real situation, so that the results of this paper only prove the theoretical possibility of strategic manipulation of the mechanism. Since preliminary simulations carried out on the real data suggest that the possibility of manipulation rarely occurs, we think that an in-depth analysis of the real case would be convenient to understand how frequent and severe are the possibilities of such manipulations.

Acknowledgements We thank the Transplant Unit of the "Ospedale S. Martino" in Genova, the Nord Italia Transplant Program and the Italian National Transplant Centre.

Appendix

Indirect mechanisms in the incomplete information setting: an impossibility result

As briefly described in Section 2, a key result in the context of mechanism design is the revelation principle. In this section we show how this can be applied in order to prove an impossibility theorem as a consequence of our examples. In the paper we have analyzed incentive compatibility of a direct mechanism, i.e. a mechanism that simply asks to each patient to report his type. In this Appendix we enlarge the set of available mechanisms, taking into account also indirect mechanisms. We therefore need some more terminology.

According to Definition 5, we consider a kidney exchange problem $K = (N, \{D_i\}_{i\in N}, \mathcal{T}^n, P, w, q)$. An outcome of the problem is an admissible matching and as we have already seen in Section 5, each patient has a preference on the set of possible matchings, depending obviously on the set of types of the involved patients. We model these preferences by saying that each patient has a utility function $u_i : \mathcal{T}^n \times \mathcal{M} \to [0, +\infty)$ (clearly we consider the expected utility when we deal with $L(\mathcal{M})$ instead of $\mathcal{M}$). We assume that the utility function is common knowledge, but not the patient's types.

Definition 6 Given a kidney exchange problem K, we call $MW(K)$ the set of maximum weight matchings corresponding to K, and we denote by l its cardinality. The function f that associates to each kidney exchange problem the element in $L(\mathcal{M})$ assigning probability $1/l$ to each matching belonging to $MW(K)$ is called maximum weight choice function.

As stated in Section 3 in a slightly different context, a direct mechanism consists of an outcome selection function $h : \mathcal{T}^n \to L(\mathcal{M})$. In order to introduce the class of indirect mechanisms, we recall the definition of a Bayesian game with consistent beliefs and Bayes-Nash equilibria (see [10]).

Definition 7 A Bayesian game is given by $(N, A, \mathbb{T}, q, (v_i)_{i\in N})$ where $N := \{1, \ldots, n\}$ is a set of players, $A := \prod_{i\in N} A_i$ is a collection of sets of actions and $\mathbb{T} = \prod_{i\in N} T_i$, with a probability distribution q over it, with T_i the set of types of each player i, and a utility function $v_i : \mathbb{T} \times A \to \mathbb{R}$. A pure strategy in a Bayesian game is a mapping from types to actions, i.e. $s_i : T_i \to A_i$.

Definition 8 The profile of strategies $(s_1^*, \ldots, s_n^*)$ is a Bayes-Nash equilibrium if for every agent i, for every type $t_i \in T_i$, and every alternative strategy $s_i \in A_i$, we have

$$\sum_i q(t_{-i}|t_i) v_i \left(t, s_i^*(t_i), s_{-i}^*(t_{-i})\right) \geq \sum_i q(t_{-i}|t_i) v_i \left(t, s_i(t_i), s_{-i}^*(t_{-i})\right).$$

(We recall that the vector v_{-i} is obtained from the vector v by deleting the i-th component).

Definition 9 A mechanism for the kidney exchange problem in the incomplete information case $K = (N, \{D_i\}_{i\in N}, \mathcal{T}^n, P, w, q)$ is given by a triple (A, o, s) where:

- $A = \prod_{i\in N} A_i$ is a collection of sets of actions;
- $o : A \to L(\mathcal{M})$ is an outcome function;
- $s_i : \mathcal{T} \to A_i$ is a set of strategies.

Definition 10 A mechanism implements the maximum weight rule f for a kidney exchange problem K in Bayes-Nash equilibrium if there is a Bayes-Nash equilibrium $s^* = (s_1^*, \ldots, s_n^*)$ of the Bayesian game $(N, A, \mathcal{T}^n, q, (u_i \circ o)_{i\in N})$ such that for all $(t_1, \ldots, t_n) \in \mathcal{T}^n$ it holds $o(s^*(t_1, \ldots, t_n)) = f(t_1, \ldots, t_n)$.

We recall the fundamental result that we have cited several times, the so called revelation principle, restricted to our case (see [1], Ch. 7, p.174).

Theorem 1 *Suppose that there is a mechanism that implements the maximum weight social choice rule f in Bayes-Nash equilibrium. Then there exists a Bayes-Nash equilibrium incentive compatible direct mechanism that also implements f in Bayes-Nash equilibrium (with the truth-telling equilibrium).*

In the rest of the paper we have shown that there exist some cases in which the direct mechanism does not truthfully implement the maximum weight social choice function. More precisely, in Example 2, we have studied such a case. As a consequence of the revelation principle, this implies that for the problem in Example 2 the maximum weight social choice rule is not Bayes-Nash implementable through any mechanism. More formally, we can collect these considerations in the following theorem.

Theorem 2 *When there are at least three patients, there exist a set of possible types $\mathcal{T}$ and a probability distribution on $\mathcal{T}$ such that no mechanism exists that Bayes-Nash implements the maximum weight social rule for every realization of the kidney exchange problem.*

References

1. Conitzer V (2006) Computational aspects of preference aggregation. Ph.D. Dissertation, Computer Science Department, Carnegie Mellon University, Pittsburgh
2. Delmonico FL, Morrissey PE, Lipkowitz GS, Stoff JS, Himmelfarb J, Harmon W, Pavlakis M, Mah H, Goguen J, Luskin R, Milford E, Basadonna G, Chobanian M, Bouthot B, Lorber M, Rohrer RJ (2004) Donor kidney exchanges. Am J Transplant 10(4):1628–1634
3. Keizer KM, de Klerk M, Haase-Kromwijk BJJM, Weimar W (2005) The Dutch algorithm for allocation in living donor kidney exchange. Transplant Proc 37:589–591
4. Kranenburg LW, Visak T, Weimar W et al (2004) Starting a crossover kidney transplantation program in the Netherlands: ethical and psychological considerations. Transplantation 78(2):194–197
5. Lucan M, Rotari P, Neculoiu D, Iacob G (2003) Kidney exchange program: a viable alternative in countries with low rate of cadaver harvesting. Transplant Proc 35:933–934
6. Montgomery RA, Zachary AA, Ratner LE, Segev DL, Hiller JM, Houp J, Cooper M, Kavoussi L, Jarrett T, Burdick J, Maley WR, Melancon JK, Kozlowski T, Simpkins CE, Phillips M, Desai A, Collins V, Reeb B, Kraus E, Rabb H, Leffell MS, Warren DS (2005) Clinical results from transplanting incompatible live kidney donor-recipient pairs using kidney paired donation. J Am Med Assoc 294:1655–1663
7. Myerson R (1979) Incentive compatibility and the bargaining problem. Econometrica 47:61–73
8. Myerson R (1983) Mechanism design by an informed principal. Econometrica 52:461–487
9. Myerson R, Satterthwaite M (1983) Efficient mechanisms for bilateral trading. J Econ Theory 28:265–281
10. Myerson RB (1991) Game theory. Analysis of conflict. (Harvard University Press, Cambridge)

11. Opelz G (1997) Impact of HLA compatibility on survival of kidney transplants from unrelated live donors. Transplantation 64(10):1473–1475
12. Opelz G (1998) HLA compatibility and kidney grafts from unrelated live donors. Transplant Proc 30(3):704–705
13. Park K, Moon JI, Kim SI, Kim YS (1999) Exchange donor program in kidney transplantation. Transplantation 67: 336–338
14. Ross LF, Woodle ES (2000) Ethical issues in increasing living kidney donations by expanding kidney paired exchange programs. Transplantation 69:1539–1543
15. Roth AE (1989) Two-sided matching with incomplete information about others' preferences. Games Econom Behav 1(2):191–209
16. Roth AE, Sönmez T, Utku Ünver M (2004) Kidney exchange. Q J Econ 119(2):457–488
17. Roth AE, Sönmez T, Utku Ünver M (2005) Pairwise kidney exchange. J Econ Theory 125(2):151–188
18. Serrano R (2004) The theory of implementation of social choice rules. SIAM Rev 46(3):377–414 (electronic)
19. Zenios SA, Woodle ES, Ross LF (2001) Primum non nocere: avoiding harm to vulnerable wait list candidates in an indirect kidney exchange. Transplantation 72(4):648–654