Analyse & Kritik 23/2001 (© Lucius & Lucius, Stuttgart) p. 262–270

Marlies Ahlert

Patterns of Decision Making in Kidney Allocation Problems

Abstract: Experts in the field of organ transplantation had to rank order a set of 32 patients according to their priority in receiving a donated kidney. The patients were described by the five characteristics that are incorporated in the kidney allocation algorithm applied by Eurotransplant. The priority rankings as defined by the experts were analyzed and patterns of decision making identified in the rankings investigated in this study. All patterns could be explained by some type of lexicographical ranking. The larger group of experts preordered tissue compatibility or, more technically speaking, the criterion of HLA match, while the complementary group applied the criterion of the length of waiting time first. Analyzing the finer decision structures of expert rankings and comparing the method of pattern exploration with a conjoint measurement analysis led to two follow-up questions: First, how can the value judgments of the experts be described adequately? Second, which type of aggregated ordering derived from the individual rankings represents them 'best'?

1. Introduction

The allocation algorithm applied in the kidney allocation process by Eurotransplant is a 'weighted point system'. The point values that define the priority of a patient on the waiting list are derived from five criteria, namely the number of HLA mismatches, HLA matchability (or chance of a very good match within one year), waiting time, distance between donor and recipient, and international exchange balance. (For a precise characterization of the allocation system, see De Meester/Persijn/Wujciak/Opelz/Vanrenterghem 1998.) Is there a justification for using a weighted sum as an allocation mechanism and, more specifically, for the choice of weights such that both choices could possibly be founded on the judgments and decisions that experts in the field of organ allocation would make?

To analyze the decision structures of experts we use the empirical data of priority rankings of hypothetical patients generated by Diederich (2001). In contrast to the conjoint measurement approach, it is not assumed here that only a weighted point system can represent the underlying decision structures of each expert. While the alternative approach simultaneously estimates weights for a weighted sum point scale that represent an aggregated form of decision making for the whole group of experts, the present one generates decision patterns that do not rely on weights. It is shown that nearly all investigated priority rankings of patients can be reconstructed by a priority ranking of the patients' characteristics.

In section 2 we present our method of investigation and discuss some formal properties of the patterns that will be focused on in section 3. In section 4 some of these considerations are applied to the results of a conjoint measurement approach and lead to the proposal of a reinterpretation. In section 5 the categories for the decision structures are developed and the observed experts' rankings assigned to them. Section 6 addresses the open social choice problem of generating a centralized rule that would be a 'good' compromise between different types of individual rules which nevertheless exhibit certain similarities.

2. Method

As already mentioned, the following analysis is based on the data set generated by the intriguing study that Diederich conducted in an effort to represent expert judgments by means of a conjoint measurement analysis of their decision behavior (see the contribution by Diederich in this volume). A set of 32 cards each containing relevant data of a patient was given to physicians who are experts in the field of organ transplantation. On each card a certain type of patient was described in terms of the following 5 characteristics (see Appendix B in Diederich 2001):

- the number of HLA mismatches (MM) of the patient's antigen groups for a given hypothetical donor kidney (numbers were chosen from 0, 1, 3, or 5 MM)
- the level of HLA matchability (low, medium, or high)
- the waiting time in years (1/2, 2, 4, or 6 years)
- the distance between the location of the donor organ and the recipient's transplantation center (no distance=same center, short, medium, or long)
- willingness of the public in the patient's country to donate kidneys within the Eurotransplant region (low, medium, or high).

The experts got instructions to order the 32 'patients' (see Appendix A in Diederich 2001). The sequence should represent the ordering in which they would assign an available kidney to the latter. For our analysis we use the data set of 16 experts' rankings as received from Diederich (see ID 01 to 16 in Appendix B, Tables 5 and 6).

Checking the axioms of Pairwise Consistency and Pairwise Separability (for definitions and a discussion of these axioms, see Ahlert/Gubernatis/Kliemt in this volume), we find very few violations of these axioms. This observation coincides with the results of Diederich's conjoint measurement analysis. The statistics show that reversals contradicting the axioms mainly occur with respect to criteria turning out to be rather unimportant to the experts. Therefore, the decision-making procedure of each expert can be reconstructed 'quite precisely' by a weighted point system applied to the set of criteria (this result from decision theory is discussed by Ahlert/Gubernatis/Kliemt in this volume, too).

In 'hard choices' subjects actually tend to decide quite incoherently. In view of the multi-attributive evaluations that play a role in organ allocation, it is therefore surprising that experts were able to order the patients in a way that caused nearly no violation of the axioms mentioned above. We could not ask the experts about the decision rules they applied, and we therefore do not know which one each expert applied according to his or her own judgment. But it is possible to identify the simplest rules that would in each instance lead to the same or 'nearly the same' ranking of the patients.

For each expert we analyze his or her ranking of the patients separately. In order to find similarities between the rankings of different experts, we consider subrankings of patients with respect to certain criteria. For instance, if all patients with 0 MM are ranked before those with 1 MM, patients with 3 MM after those with 1 MM, and patients with 5 MM last, this defines a substructure of the rankings. In the example any ordering that applies the criterion of HLA match first (lexicographically preorders it) would lead to this subranking. Subsequently, we develop and investigate several decision structures based on subrankings enabling us to characterize the patterns of the experts' decision making.

3. Weighted Sum Representations of Lexicographical Orderings

Before we go into detail regarding the observed decision structures, we would like to point out the relation between weighted point systems and lexicographical orderings. Given a set of criteria and a set of alternatives that have to be ranked, a weighted point system does the following: For each criterion, each alternative is assigned a point value that is monotonically related to the degree of fulfillment of the criterion. There is a nonnegative weight that is assigned to each criterion (independent of the alternatives). This weight factor of the criterion has to be multiplied by the point value of the alternative. The sum of these products (summation over all criteria) then defines the final point value of the alternative. The ranking of the alternatives is defined by the natural order between these weighted sums.

The weight factors can be seen as a way of handling trade-offs between different criteria according to a 'measure of relative importance'. Among these weighted point systems there are some extreme ones that avoid trade-offs between criteria for instance, if a difference in the weighted value with respect to the most important criterion can never be beaten by any difference between weighted values of less important criteria. Within a given ranking of criteria according to their relative importance, trade-offs between all criteria are avoidable by defining the weight structure in relation to the point values such that differences in the weighted point value of any single criterion can never be outweighed by any subset of less important criteria. (The decimal system of ranking numbers is an obvious example of this.)

However, under such conditions it is not necessary to develop or represent the ranking of alternatives by a weighted point system. There are no weights needed since trade-offs do not have to be made. Actually, no definition of point values for any single criterion is then needed, either. The ordering can be defined without any numerical representation. We have to assume that there is a ranking of the criteria with respect to their importance, and an ordering of the alternatives with respect to each criterion. The induced lexicographical ordering compares any two alternatives with the most important criterion first. If there is a tie, the next criterion has to be applied, and so on. If there is a tie along all dimensions, the alternatives are seen as being indifferent. In all other cases some criterion is decisive in defining which alternative is to be preferred. Though lexicographical orderings can be interpreted as special, or better, extreme cases of weighted point systems, they are of a much less demanding formal structure.

4. A Reinterpretation of Estimated Importance and Utility Scores

The simple insight that lexicographical orderings are limiting cases of weighted point systems may also shed some light on the interpretation of empirically observed or estimated weights and point systems. This holds good for the analysis of individual decision making as well as of aggregated data, like e.g. the estimated aggregated importance and utility scores of characteristics in a conjoint measurement approach. Maximally achievable weighted point values per criterion (especially of criteria with low importance) have to be compared to the feasible steps in the weighted point structure measuring any other, more important criterion. If there is a single criterion or even a subset of criteria without the impact of compensating for even the smallest possible change in another criterion or another set of criteria, then this fact should be represented by some lexicographical structure.

For instance, in Table 3 of Diederich's paper, the average 'utility' scores for the levels of the criteria are estimated. The data show that a difference between two adjacent levels of mismatches can never be compensated for by any other criterion alone except for waiting time. A difference of two or more levels of waiting time (in the table this would mean more than 3 1/2 or 4 years) is needed to outweigh a difference of one level in the ranking of mismatches. An analogous observation holds for the criterion of waiting time and the set of the criteria: matchability, distance of kidney transportation, and national donation balances. This means that the estimated average decision rule can be interpreted as being 'almost' lexicographical. It cannot be represented by a purely lexicographical ordering. The decisions, however, can be generated from a ranking that allows for some exceptions from the strict ordering of the criteria in cases where large improvements with respect to a less important criterion are realizable by a very minor reduction of the fulfillment of a more important criterion (for a discussion of this type of rule, see the paper by Ahlert/Kliemt in this volume).

Let us remark here, without following this line of thinking in the present paper, that these types of rules, by including some conditional aspects for exceptions, can also be seen as being modifications of very simple or basic rules. A possible pattern would be as follows: Use the basic rule, unless condition A applies. If A then do B. Different patterns can be generated in this way. These procedures which strongly recall the 'unless clauses', well known from expert systems, are less complex than a weighted point system, but would lead to the same ranking of alternatives and could equally be used to create a common 'representative' rule.

5. Structural Analysis of the Experts' Rankings

5.1 Coarser Structure

This section aims at construing a decision rule for each subject separately by looking at his or her ranking of the patients. Since several decision rules might lead to the same ranking, we concentrate on simple decision rules. We already know that weighted point systems would be applicable because the axioms of pairwise separability and pairwise consistency are 'approximately' fulfilled. As an implication of the discussion above the goal of simplicity leads to the following research question: How many decisions of experts can be explained by some ('simple') priority ordering of the patients' characteristics?

We concentrate on the criteria HLA matching and waiting time, the most important characteristics for the majority of the group of experts considered. Starting with HLA match, we check which experts rank this criterion first while considering everything else as less important. This type of decision rule leads to a division of the 32 cards into four subsets and to a ranking of these subsets ('equivalence classes'). The first subset includes all patients with 0 mismatch, the second subset those with 1 mismatch, the third subset the patients with 3 mismatches, and the last subset those with 5 mismatches. (Here we do not consider the subrankings within each subset of patients.) In the orderings of the experts with the ID 02, 04, 06, 08, 12, 14, and 15 the four subsets of patients are ranked in this way.

Observation 1:
6 out of 16 experts rank the criterion of HLA match lexicographically first and discriminate between all different numbers of mismatches.
Some other experts (ID 07, 10, 11, 13, and16) show the same first priority concerning the criteria, but modify the HLA match characteristics. They identify 0 and 1 mismatch in their evaluation and thus show the same ordering of three subsets. First, there is the group of patients with 0 or 1 mismatch, followed by those with 3 mismatches and those with 5 mismatches.

Observation 2:
5 out of 16 experts rank the criterion of HLA match first, but do not discriminate between 0 and 1 mismatch.

Three experts (ID 01, 05, and 09) show an ordering of the cards that coincides with the priority of waiting time over all other criteria. This means that they ordered the patients starting with those having spent a waiting time of 6 years, and so on, with decreasing numbers of years. Two more experts used a rule that can be explained by first considering waiting time except for cases of 0 or 1 mismatch (i.e., 'order according to length of waiting time unless ... !'). These two 'best' levels of the criterion 'number of HLA mismatches' are able to compensate for one level in waiting time.

Observation 3:
3 out of 16 experts rank the criterion of waiting time first and discriminate between all given numbers of years. 2 experts use an almost lexicographical ordering with waiting time as the first criterion.

These results coincide with Diederich's results in Table 9. For each expert, she presents the statistically determined 'importance ranking' of the five criteria. However, from the experts' rankings it cannot be concluded what type of decision rule they used. The statistical result, e.g. of ID 03, says that for this expert the number of HLA mismatches is the most important and waiting time the second most important criterion. The decisions of expert ID 03 can also be generated by applying the criterion of waiting time first, except for cases of 0 or 1 mismatch that can compensate for one level of waiting time. Of course, this ranking is not very different from one that identifies 0 and 1 mismatch and shows this as first choice except for cases of long waiting time. Which criterion is more important in this case?

The advantage of the conjoint measurement approach is that all positional rankings are taken into account simultaneously, whereas the effort to construct a rule that may have been used will always have to deal with several possible forms and accept the existence of exceptions from the rule. The latter problem could be avoided, or at least reduced, if the subjects were asked to verbally describe the procedure of their decision making. In any event, in case of a weighted sum representation as well as a rule-based representation of choice making, all one can say is that the individuals chose *as if* they were acting as represented. It is an obvious research question for a future project to scrutinize more closely which mental processes did in fact lead to the observed behavioral patterns. It is quite clear that simple rules of thumb rather than weighted sums are in fact driving the process. Whether the former are the simplest ones as postulated here, or even somewhat different rules, remains to be seen.

We note that all experts apply one criterion lexicographically (14), or at least almost lexicographically (2), first. Eleven experts decide that HLA match is most important to them, five experts choose waiting time, but two of the latter make some concession with respect to HLA match.

5.2 Finer Structure

Analyzing the next finer structure of decision making, we start with those experts whose rankings can be represented by a strictly lexicographical ordering with respect to the two criteria they consider most important. We then analyze those rankings that show some exceptions relating to the second criterion. These decisions can be modeled by including some further aspects of the remaining criteria as conditions for the exceptions. We find the following:

- first HLA match, second waiting time: 1 pure, 2 including further aspects
- first HLA match, second matchability: 1 pure, 1 including further aspects
- first HLA match, second a mixture: 1
- first HLA match identifying 0 and 1 mismatch, second waiting time: 2 pure, 3 including further aspects
- first waiting time, second HLA match: 2 pure, 3 including further aspects.

Summarizing these findings we conclude that the combination of HLA match and waiting time in some lexicographical, or almost lexicographical, structure explains 13 out of 16 rankings. The remaining 3 rankings show matchability as the second criterion after HLA match or a combination of matchability and waiting time.

Observation 4:
A large majority of the experts do not consider trade-offs between the two criteria, HLA match and waiting time (or matchability), they consider most important, on the one hand, and any other aspect, on the other.

All other criteria are used by the experts in cases where, with respect to the criteria they consider of primary importance, there is no or almost no difference between two patients. Here the ranking of importance varies, even between experts from the same type of lexicographical structure, as defined above. In the whole data set, there are no two completely identical rankings. It does not seem to be appropriate to analyze the 'fine-tuning' of the experts, since no generalization would be possible. Here the importance ranking as determined by the conjoint measurement approach (Diederich, Table 9) gives a better overview of the relations between the less important criteria.

6. Conclusions

The fact that nearly all rankings can be generated by some priority ordering of the patients' attributes does not necessarily imply that the experts had this structure in mind. On the other hand, the task of sorting a pack of cards with descriptions of patients with respect to five categories has some similarities with the task of ordering numbers in a decimal system, of putting cards with words into alphabetical order, or of arranging playing cards, all of which are examples of lexicographical orderings. Maybe the aim to rank the patients in a 'systematic' and nonarbitrary way made the participants of the study fall back on some method learned and tested in other situations. Lexicographical orderings are familiar and, whenever the ranking of the attributes is clear, easy to handle without a great risk of making mistakes. Reliance on familiar and simple lexicographical ordering procedures would also explain why there are so few violations of the consistency and separability axioms, especially with respect to the most important criteria.

It is important to repeat that, according to our analysis, there is strong evidence that the experts do not consider trade-offs between the criteria they regard as most important. A majority of them applies HLA mismatch first, while a minority goes for waiting time first. Focusing just on this result, how can we relate it to any centralistic rule like the one used by Eurotransplant? If we want to generate a 'representative' rule from the decisions, we have to decide on the structure of that rule. Should it be a lexicographical (or almost lexicographical) rule again, or a rule with some conditions added for exceptions in order to include minority opinions. Or is a rule that needs some cardinal representation inevitable?

The conjoint measurement approach offers one possibility of calculating a compromise rule from the set of rankings using a cardinal representation. Since the decision making of the experts does not need a cardinal representation to be generated, it would be preferable to work out a compromise on an ordinal basis, possibly combined with some conditional statements. This leads to the familiar but very challenging social choice problem of aggregating individual orderings (in our case of a lexicographical structure with finitely many criteria) into some collective ordering. It is not obvious which properties the aggregation procedure should have, nor which properties any resulting collective rule might have. But any comparison with the results of the Eurotransplant algorithm, and any evaluation of the established procedure in view of possible alternative collective decision procedures would of course depend on the latter aspects.

There is another problem related to the comparison of individual decision making and centralistic decisions. What type of criteria should be included in the collective rule? Would it be appropriate to take the characteristics of the patients a physician would look at, or are there more general distributional aspects that would not be taken into account by any individual decision maker, but may be important for a collective rule (e.g. input-output balance between countries). Any answer to this question presupposes the existence of some quality measurement for each type of decisions. Experts will probably try to make 'good' decisions on the level of individual patients. Their focus is on the set of patients they are confronted with (in reality or in the experimental situation). As a consequence, they may focus on personal characteristics of the patients rather than statistical data. A collective mechanism may keep to this level, but may also include some statistical aspects in order to cover criteria of sufficient supply or of distributive justice between subgroups (like input-output balance). Taking nonpersonal characteristics of the patients into account may improve the quality of the collective rule in terms of more efficiency or equity, whatever the aim or consequences of the additional criteria are. Therefore, we believe that comparing a rule that is aggregated from individual decisions and a rule that is designed as a collective rule must not neglect the fact that the quality standards of individual and collective decision making may be different. This problem in a way echoes the old insight that there are unintended consequences of human action that are not the result of human design. But the problem will take a somewhat new turn in the context of organ allocation where the fixing of collective allocation rules often seems to have been completely dictated by considerations of statistical outcomes on the collective level. In future research on organ allocation this problem, too, needs to be modeled and discussed in more detail.

Bibliography

Ahlert, M./G. Gubernatis/H. Kliemt (2001), *Kidney Allocation in Eurotransplant. A Systematic Account of the Wujciak-Opelz Algorithm*, in this volume

—/H. Kliemt (2001), *A Lexicographical Decision Rule With Tolerances. The Example of Rule Choice in Organ Allocation*, in this volume

De Meesters, J./G. G. Persijn/T. Wujciak/G. Opelz/Y. Vanrenterghem (1998), The New Eurotransplant Kidney Allocation System, in: *Transplantation 66*, 1154–1159

Diederich, A. (2001), *A Rational Reconstruction of Expert Judgements in Organ Allocation. A Conjoint Measurement Approach*, in this volume