Matching Models and Optimal Registry for Voluntary Organ Donation Registries[1]

F. Fève[2] and J.P. Florens[3]

(Preliminary and Incomplete)

May 2005

[1] This work was supported by the European contract MADO No: QLG7-2001-00065) and EFG (Etablissement Français des Greffes). The authors thank Anne Cambon-Thomsen (INSERM U558, Toulouse) for initially raising the question, FGM (France Greffe de Moelle) and the French histocompatibility laboratories for providing HLA data, CNG (Centre National de Genotypage, Evry) for the microsatellite data, Pierre-Antoine Gourraud (INSERM U558) for checking and formatting the data. They are grateful of Fabrice Collard, Pierre Dubois and Jean–François Eliaou for helpful discussions and comments.
Address of the authors: IDEI, Université Toulouse 1 Sciences Sociales, Manufacture des Tabacs - Bât F, 31000 Toulouse, France, Tel: +33(0)5.61.12.85.90, Fax: +33(0)5.61.12.86.37, email: feve@cict.fr, florens@cict.fr

[2] University of Toulouse (GREMAQ and INSERM U558)

[3] University of Toulouse and IUF (GREMAQ)

Abstract

This paper considers a general decision model of voluntary organ donation registry. The main example is given by the marrow donors registries organized in most of the major countries. A registry is a list of voluntary donors with known type and transplantation requires identical type between donor and receiver. As typing has a relative high cost registry should be organized in an optimal way in order to increase the probability for a patient to find a donor. This paper shows what is the optimal (but not implementable) registry and how filtering mechanism may be used in order to improve actual registry.

Keywords: Decision theory, evaluation of information system, marrow donors registry.

JEL Classification: I18, C70, C61.

1 Introduction

The purpose of this paper is to perform an evaluation of a mechanism for the constitution of organ donor registries. The aim is to increase the adequacy between the file of potential volunteer bone marrow donors and the needs of patients. Alive voluntary organ donors are needed in order to treat some diseases and to perform grafts. These organs are taken just in time before transplantation. It's impossible to preserve them. One example is typically the bone marrow transplantation (CSH) aimed to treat blood diseases especially leukemias and immune-deficiencies diseases. A graft from a donor to a patient is possible only a compatibility condition. This condition is, in theory, the identity of the HLA system where the HLA (Human Leucocytes Antigens) is characterized by a double sequence of alleles of a set of genes on the pair of the 6th chromosome; A simplified view of the HLA consists by considering only three genes, A, B and DR and the type of an individual is described for example by (1,2) (2,44) (3,4) which means that the pair of gene A is 1 and 2, of gene B 2 and 44 and of gene C is 3 and 4. Each pair is ordered because we cannot observe on which chromosome the alleles are. It should be underline that, contrarily to many problems in economics, individuals do not know their own type. Moreover the typing of an individual has a substantial cost. As the number of possible alleles for each gene varies between 10 and 40, the number of theoretical possible HLA is huge (several millions). The number of possible types and their inequal distribution implies that a large number of potential volunteer bone marrow donor is needed. The French Registry contains approximatively 120 000 donors and is interconnected with the worldwide file.The total number of potential donors in the world is more than 6 millions but only a small proportion of patients found an available compatible donor. In France this proportion is smaller than 10% a compatible donor [1]. Moreover it has been empirically verified that the efficiency of the registries (in term of proportion of patients who find a donor) increases very slowly when the size of the registry increases (see Martha...). The aim of this paper is to formalize this problem and to analyse the problem of how to improve the organisation of this kind of registries. Basic definitions are given in section 2. A theoretical concept of optimal registry is conducted in section 3. Then we consider implementable improvement of the registry based on a filtering system where a low cost information associated to the type may be obtained. The optimal use of an information and the selection of optimal information are theoretically computable. The problem is practically untractable due to its dimensionality.

[1] We eliminate patients for who a family donor may be found

We propose in section 4 a feasible strategy based on simulations.

2 Donors, receivers and registry

Each individual of the population is characterized by a type j belonging to a finite set $1, ..., J$. For example the type is the HLA phenotype, i.e. the list of the alleles on the two chromosomes of loci A,B and DR registered with a given precision ("two digits" or "four digits" in the HLA case). The receivers are drawn in the population and the frequencies of types for receivers is described by a probability vector p_j $(p_j > 0 \sum_{j=1}^{J} p_j = 1)$. By construction of the list of the types all the p_j are strictly positive. A registry is a list of donors recorded by their identity and their type. Two conditions are necessary in order to do a transplant to a given receiver :

- The existence in the registry of donors of individual of the same type. We assume that only perfect matching transplants are realized and we never introduce an idea of distance or almost compatibility between donors and receivers.

- At least one compatible donor should accept the transplant (or should be availlable for the transplant). Multiple causes exist for non availibility (pregnancy, professional requirements, illness...). We summarize this complex phenomena by assuming that a compatible donor "accept" the transplant with a probability a. Acceptation decisions for different donors are independent events.

A registry design is defined by two components :

- an initial registry : this initial registry is characterized by its size N_0 and by the number $N_0 j$ of donors of type j. This number may be equal to 0 for many types.

- an increment process defined by the number N of new donors introduced in the registry and by a sampling mechanism of the types described by a vector q_j $(q_j \geq 0 \sum_{j=1}^{J} q_j = 1)$ of frequencies. For example if donors and receivers are drawn randomly in the same population $p_j = q_j$ and are equal to the frequency of type j in the population.

We should underline that individuals and registry management ignore the types. Typing is a complex operation only realized where a new donor is introduced in the registry. Then, in practise, the registry management cannot choose q_j. However we first imagine situation where $(q_j)_j$ can be selected arbitrarily by the registry management ("first best" approach) and we consider secondly implementable choices of q_j ("second" best approach). In all own analysis N is given. In practice N may be constrained by the arrival process of new potential donors and by the budget of the registry organization because introducing a new donor is costly.

We consider in this paper only a single period model : starting for an initial registry we consider its improvement by a unique increment and not by several increments through multiple periods. The mathematical tools to model and to solve this dynamic version used more technicalities are needed to solve this dynamic programming problem. A dynamic model requires more abstract mathematical tools belonging to the theory of dynamic programming problems.

As an illustration consider the french registry of bone marrow donors. The types are defined by HLA haplotypes A,B,DR recorded in two digits. The current registry contains approximatively 100 000 donors and an increment of 10000 by year is scheduled. If we want to analyze the one year mechanism N is 10 000 but we may also consider a long term variation ($N = 100\ 000$ or more). The number J of possible types is a crucial element of the analysis and will be discussed later on. More than 60 000 types are present in the registry. At the world level the interconnection between the national registries gives a total registry of more than 6 millions which also increases by several hundred of thousands people each year. More than 400 000 types have been observed.[2] A registry system is the defined by J, the $p_j' s$, the $N_{0j}' s$, a, N and the $q_j' s$. Such a system may be evaluated.

We propose to evaluate a registry by the expected probability to not find a donor for a receiver. To illustrate this concept consider just the simple case where $N_0 = 0$ (no stock) and $a = 1$ (all compatible donors accept the transplant). For any receiver the non-realization of a transplant (all the donors have a type different of j) is a random event which has a probability of $(1 - q_j)^N$ if the type of the receiver is j. As the type of the future receivers are not given when the registry is designed we consider the expectation of this probabilities through the different types :

$$L = \sum_{j=1}^{J} p_j (1 - q_j)^N$$

This quantity may be viewed as the evaluation of the registry system and 1-L is equal to the probability of any receiver to find a donor.

[2] The list of possible alleles on each locus A,B, DR is probably known and then defines a huge list of potential types. However most of associations don't exist and the number of sequences A,B, DR is smaller than the product of the number of alleles on each locus.

Remark 1: An alternative criterium for the evaluation of the registry design would be based on the expected waiting time of a patient. Let us assume that N new donors are drawn with a probabilities $(q_j)_{j=1,\ldots,J}$. For a patient where type is j, the waiting time is 0 with probability $1-(1-q_j)^N, 1$ with a probability of $(1-q_j)^N(1-(1-q_j)^N)$. In general the waiting time is t with a probability of $(1-q_j)^{Nt}(1-(1-q_j)^N)$. The expected waiting time of this Pascal distribution is equal to $\frac{1}{1-(1-q_j)^N}$. We average this expected time with respect to the patient's type and we get the following evaluation criterium:

$$L_1 = \sum_{j=1}^{J} p_j \frac{1}{1-(1-q_j)^N}$$

Remark 2: In the previous definitions we consider the cases where a patient does not find a donor only. We may also take into account cases where a donor is found. For example if A is the cost of not find a donor and B the value where a donor is present in the registry, the evaluation of the registry design is :

$$L_2 = \sum_{j=1}^{J} p_j(A(1-q_j)^N - B(1-(1-q_j)^N) + B\sum_{j=1}^{J} p_j$$

It may be very easily shown that optimal designs of registries based on L_1 or L_2 instead of L will give same results. Then we keep L as an evaluation criterium but we extend its evaluation to case where $N_0 \neq 0$ and $a \neq 1$.

Proposition 1: If N is large and the q_j are small the evaluation of a registry system measured by the expected probability is not found a donor may be approximated by :

$$L = \sum_{j=1}^{J} p_j (1-a)^{N_{0j}} e^{-aNq_j}$$

Proof : let fix j the type of a receiver. The number of donors of this type is $N_{0j}+m_j$ where m_j is drawn by a Binomial distribution :

$$Prob(m_j) = C_N^{m_j} q_j^{m_j} (1-q_j)^{N-m_j}.$$

Given j and m_j the probability of to not find a donor is $(1-a)^{N_{0j}+m_j}$. Then given j only, the probability to not find a donor is

$$\sum_{m_j=0}^{N} (1-a)^{N_{0j}+m_j} C_N^{m_j} q_j^{m_j} (1-q_j)^{N-m_j}.$$

Then

$$\begin{aligned} L &= \sum_{j=1}^{J} p_j \sum_{m_j=0}^{N} (1-a)^{N_{0j}+m_j} C_N^{m_j} q_j^{m_j} (1-q_j)^{N-m_j} \\ &= \sum_{j=1}^{J} p_j (1-a)^{N_{0j}} E[(1-a)^{m_j}] \end{aligned}$$

where m_j is drawn by the binomial distribution. For large N and small q_j it is wellknown that this Binomial distribution is approximatively a Poisson distribution parametrized by $\lambda_j = Nq_j$. Moreover an elementary computation shows that $X \sim \Im(\lambda)$ implies $E(b^X) = e^{\lambda(b-1)}$. Then :

$$L = \sum_{j=1}^{J} p_j (1-a)^{N_{0j}} e^{-aNq_j}$$

■

In particular if the potential donors arrive randomly the probability to not find a donor is :

$$L = \sum_{j=1}^{J} p_j (1-a)^{N_{0j} l^{-aNp_j}}$$

The result given in proposition 1 will be useful to characterise the efficiency of a design (q_j) or to compare several designs.

3 Optimal registry: theory

In this section we assume that the institution which has in charge the management of the registry optimizes the registry design in order to minimize the evaluation criterium under a budget constraint. This institution has a fixed total budget B and it is assumed that the cost of the registry is linear (any donor costs b). The budget constraint reduces in that case to the elementary relation :

$$B = bN$$

for which N is determined and equal to B/b. A more sophisticated cost function $C(N)$ may be introduced and the constraint becomes $B = C(N)$ but in all cases the number of new donors N follows from the budget constraint and this number N is not a random element. The main element of this assumption is that the cost function of the treatment of donors (which contains essentially the typing cost) cannot be influenced by the institution which manages the registry. This assumption may be false in practice : registry management and typing laboratories are often both controlled by the public health administration which may be modified the cost function. In this section we only consider the case where N is determined by the budget constraint.

The problem then reduces to minimize the evaluation criterium with respect to the drawing design of the donors $(q_j)_{j=1,...,J}$. This analysis has only a theoretical objective because the result will require to be implemented the knowledge of the types. The result has however an interest as a reference theoretical optimal registry.

Equivalently the problem is to minimize

$$L = \sum_{j=1}^{J} p_j (1-a)^{N_{0j}} e^{-aNq_j}$$

with respect to the q_j's under the constraints:

$$\sum_{j=1}^{J} q_j = 1 \text{ and } q_j \geq 0 \ \forall j = 1, ..., J.$$

where the p_j, a, the N_{0j} and N are given.

Let us denote by $\tilde{q}_j$ and $\tilde{L}$ the solution of this problem.

This optimization problem has no solution in a closer form and should be performed numerically. Theorically this optimisation is not difficult even if the objective is a non linear function of q_j's. The constraint are linear : one an exact constraint and J are inequality constraints. However the problem is almost untractable in practice because the dimension J of the q vector is extremely large. We will show later on some example of this computation in models with "small" J (1 around 1000).

One can remark that the minimization of L under the equality constraint only has an elegant solution which provides a bound of the efficiency the registry.

Proposition 2 : The minimum of L with respect to the q_j under $\sum_{j=1}^{J} q_j = 1$ is reached for

$$q_j^0 = \frac{1}{J} + \frac{1}{aN}\left\{lnp_j - \frac{1}{J}\sum_{\ell=1}^{J} lnp_\ell\right\} + \frac{ln(1-a)}{aN}\left\{N_{0j} - \frac{N_0}{J}\right\}$$

and the optimal value of L is equal to

$$L^0 = J\bar{p}(1-a)^{\frac{N_0}{J}} e^{-\frac{aN}{J}}$$

where $ln\bar{p} = \frac{1}{J}\sum_{j=1}^{J} lnp_j$

<u>Proof</u>: We replace q_J by $1-\sum_{j=1}^{J-1} q_j$ and we compute the first order condition of the minimization:

$$\frac{\partial L}{\partial q_j} = p_j(1-a)^{N_{0j}} aNe^{-aNq_j} + p_J(1-a)^{N_{0J}} aNe^{-aNq_J} = 0 \qquad \forall j = 1, ..., J-1.$$

Then:

$$p_j(1-a)^{N_{0j}} e^{-aNq_j} = \text{ constant and } \sum_{j=1}^{J} q_j = 1$$

$$\Rightarrow q_j^0 = \frac{1}{aN}\{lnp_j + N_{0J}ln(1-a)\} + C$$

Using $\sum_{j=1}^{J} q_j = 1$ we get

$$C = \frac{1}{J} - \frac{1}{aN}\left\{\frac{1}{J}\sum_{\ell=1}^{J} lnp_\ell + \frac{N_0}{J}ln(1-a)\right\}$$

from which the q_j^0's are derived.

The solution of the first order conditions is a minimum because the function L is convex as a function defined on the q_j's.

The value of L^0 is immediately obtained by replacing q_j by q_j^0.

■

These results are easy to interpret and require several comments.

Remark 1: The value L^0 is obtained by relaxing some constraint satisfied by $\tilde{L}$. As a consequence :

$$L^0 \leq \tilde{L}$$

or

$$1 - L^0 \geq 1 - \tilde{L}$$

The value $1 - L^0$ then gives an upper bound to the probability to find a donor and can be view as an (optimistic) measurement of the maximal efficiency of a registry.

Remark 2: This upper bound $1 - L^0$ depends in a few number of characteristics of the registry system: It depend only:

- on the sizes of the initial and the incremental registries N_0 and N.
- on the probability a to fo a donor to be available for a transplant
- on the number of types J
- on a characteristic of the dispersion of the distribution of the types in the receiver's population. This characteristic is the geometrical mean $\bar{p}$.

More precisely the key elements of the efficiency of the registry are $J\bar{p}$ (ratio the geometrical mean $\bar{p}$ and of the arithmetical mean $1/J$) and the relative sizes of the registries relatively to J ($\frac{N_0}{J}$ and $\frac{N}{J}$). Finally the parameter a is a key element of this efficiency.

Remark 3: The q_j^0 depends on three components:

- The uniform distribution $\frac{1}{J}$
- A measure of the importance of p_j with respect to $\bar{p}:$ types j for which p_j is greater than $\bar{p}$ should have q_j^0 greater than $\frac{1}{J}$.
- A measure of the importance of N_{0j} relative to the mean size of each type the initial registry ($\frac{N_0}{J}$). This measure is weighted by $ln(1-a)$ which is negative. Then over represented types in the initial registry have a q_j^0 inferior to $\frac{1}{J}$.

Unfortunately direct application of this formulae may lead to negative values if J os large and N relatively small.

For Large N the optimal q_j^0 converges to $\frac{1}{J}$ (the uniform distribution) and are then positive. However we will see in the next section that the value of N for which $\frac{1}{J}$ may be accepted is extremely large if J is also large.

Remark 4: Given a populations characterises by the number of the types J, their frequencies p_j, the number of patients P (during a sufficiently long period of time) it is possible to evaluate the optimal size of the registry. This size depends on the value of providing an available compatible donor to any patient. This value may be evaluated using the expected efficiency of a graft. We assume that the registry design is optimal and the social benefit of a registry of size N is:

$$VP(1 - J\bar{p}e^{-\frac{aN}{J}}) - bN$$

In that expression we have considered an optimal registry constructed in one period and not by improvement of an initial registry ($N_0 = 0$). The first component of this expression represents the value of the registry (number of patient × probability to find a donor × value of the matching) and the second turn is the total cost.

This function has always a unique maximum for the value:

$$N = \frac{J}{a} ln \frac{VP\bar{p}a}{C}$$

However this result is relevant only if this value is positive are equivalently if

$$\frac{VP\bar{p}a}{C} > 1$$

We have previously remarked that the element parameter of the population is hot $\bar{p}$ but $J\bar{p}$. Then this condition gives an upper bound for the number of type :

$$J < \frac{VPaJ\bar{p}}{C}$$

Equivalently if J is given this equality may be view as defining a threshold on C (given V) or V (given C).

Remark 5: Even if the FGM file is composed by a large number of donors some selection bias exist in this sample. The number of types in reality is certainly much higher than the number of types observed in the FGM file and numerous rare types are certainly not observed in this survey.

The mado file was designed for testing the relation between HLA types and microsatellites information (see later). This sample contains multiple selection bias. The frequency of the observed types are biased and has been redressed. Moreover there are also bias in the selection of types present in this sample. This explains the difference in the $J\bar{p}$ of the mado sample and of FGM file.

Remark 6: The size of the initial registry and of its increments relatively to the number of types has been calibrated in order to reproduce the observed efficiency of the registries.

4 Optimal registry: Some simulations

In order to shed light on the previous mathematical results and to get more intuition on their contain we have done several simulations based on specific models of the registry system. These simulation are done using the following principles: we consider a list of types and their frequencies, a value of a, a size N_0 of the initial registry and of values of N_{0j}. Finally different sizes N of the incremental registry are considered. Essentially the objective is to compare the value L if donors arrives ate the frequencies p_j (no selection), "optimistic" optimal L^0 and in some cases real optimal value of the registry $\tilde{L}$.

The different models considered are all based on the french registry (France Greffe de Moelle registry) hereafter FGM). In order to calibrate our simulation we should underline the following result. In 2003, the registry had 120 937 donors[3] and 62 french receivers under 813 found an available donor. The current value of our critera L is then equal to 0,92.

We have exact three examples of a list of types from FGM and we then have three values of J : 1162 , 4648 and 66 164. The last one correspond to the list of observed type in France provided with their frequencies (these frequencies are derived from a file of 107 925 donors provided by FGM). In our simulation we have assume that this list represents the whole list of types. This is false in reality but this define the model. The sample of 4648 was the "Mado sample" and the first one is 1/4 of the mado sample. If $a = \frac{1}{3}$ the condition of the simulation are summarized in the following table:

Table 1

Model	J	N_0	$J\bar{p}$
1	1 162	255	0.48
2	4 648	1 000	0.49
3	66 164	11 952	0.78

N = 0,10%, 25%, 50%, 100% of N_0

1: A Mado quarter (allows most of numeric computations)
2: Mado
3: FGM

[3] Data are given on the web site of FGM (www.fgm.fr)

For the first model the "small" size of the number of types allows a numerical computation "of the $\tilde{q}_j$ and of $\tilde{L}$"[4]

Results are summarized in table 2 and graphs of the $\hat{q}_j$ are given in graphs 1 and 2.

Results of model 1
$J = 1\ 162$
$N_0 = 255$
a =1/3

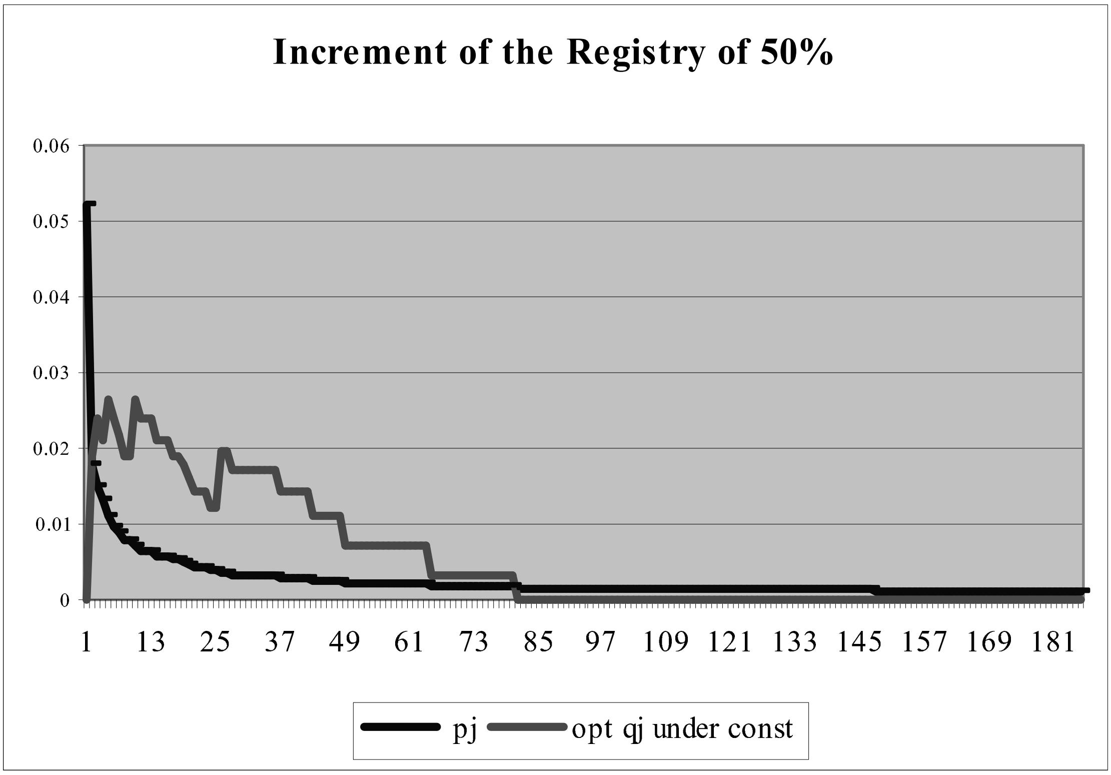
Table 2
Expected probability to not find a donor
MADO subsample

N	N_0 Selection	Optimal implementable	Optimal
$10\% N_0$	0.61	0.55	0.43
$25\% N_0$	0.59	0.54	0.43
$50\% N_0$	0.57	0.51	0.42
N_0	0.53	0.47	0.41
$2N_0$	0.47	0.42	0.38

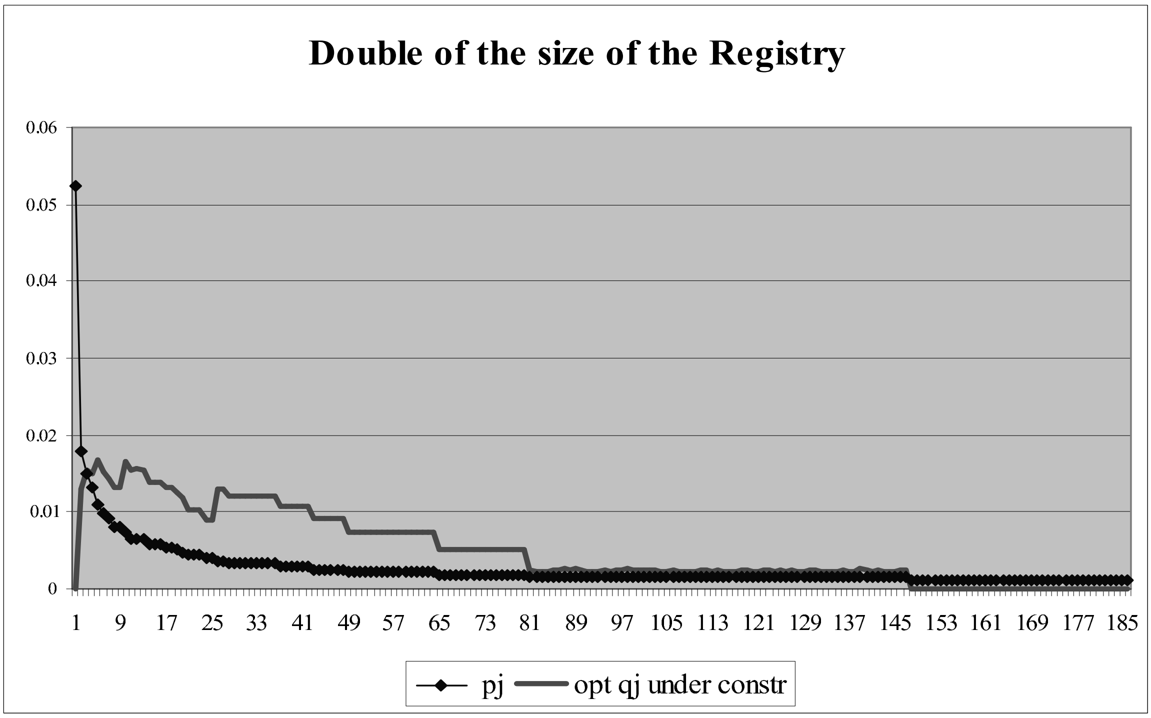
[4] Computation has been done using Matlab...

Graph 1

$N = 127$


Increment of the Registry of 50%
0.06
0.05
0.04
0.03
0.02
0.01
0
1
13
25
37
49
61
73
85
97
109
121
133
145
157
169
181
pj
opt qj under const

Graph 2
$N = 255$


Double of the size of the Registry
0.06
0.05
0.04
0.03
0.02
0.01
0
1
9
17
25
33
41
49
57
65
73
81
89
97
105
113
121
129
137
145
153
161
169
177
185
pj
opt qj under constr

For the model 2 we have to compare L with no selection, L_0 and L with some particular non optimal q_j (for instance frequent types in the initial registry have been eliminated with three definitions of frequent : more than 3, 4 or 5 *** in the initial registry. We have also eliminated frequent and rare types).

Results of model 2
$J = 4\ 648$
$N_0 = 1\ 000$
a =1/3   823 types

Expected probability to not find a donor
MADO sample

Elimination based on the initial registry									
N	$a = 1/3$			$a = 1/6$			$a = 1/2$		
	N_0 selection	optimal	optimal implementable	N_0 selection	optimal	Optimal implementable	N_0 selection	optimal	optimal implementable
0	0.637	x	x	0.7732	x	x	0.5388	x	x
100	0.6253	0.45	0.6175	0.7624	0.47	0.7619	0.5283	0.42	0.5139
250	0.6088	0.44	0.5874	0.747	0.47	0.7431	0.5137	0.41	0.4768
500	0.5841	0.44	0.5404	0.7286	0.46	0.7128	0.4917	0.40	0.4208
1000	0.5425	0.42	0.4575	0.6831	0.46	0.6558	0.4546	0.38	0.3277

Elimination based on the initial registry						
N	N_0 Selection	Optimal	$n_{oj} \geq 3$	$n_{0j} \geq 4$	$n_{oj} \geq 5$	$n_{oj} \geq 5$ et $N_{0j} = 0$
			0,1263	0,0826	0,041	0,553
0	0,75	x	x	x	x	x
100	0,74	0,641	0,737	0,736	0,735	0,731
500	0,7	0,623	0,706	0,703	0,697	0,687
1000	0,666	0,601	0,67	0,6666	0,657	0,648
1500	0,634	0,58	0,638	0,633	0,622	0,619
2000	0,606	0,559	0,609	0,604	0,592	0,597

Finally we have made a simulation at a national level. For example consider a case where 1 200 000 types exist and $J\bar{p} = 0.85$. If $a = \frac{1}{6}$ the optimal efficiency of a registry of 1 200 000 locus is 0.84 (to compare to 0.92 observed in France). If the registry increases to 240 000 (resp. 1 000 000) the optimal probability to not find a donor decreases to 0.82 (resp 0.74).

Results of model 3
$J = 66\ 164$
$N_0 = 11\ 952$
$a = 1/3$

Expected probability to not find a donor

	N_0 selection	Optimal selection (non implementable)
0	0.777	0.726
10 %	0.771	0.721
25 %	0.763	0.715
50 %	0.749	0.704
100 %	0.723	0.683

The main conclusions of these simulations are the following:

1. An important increment of the donors files (multiplication by 2 or 3 the number of donors) has a relatively low impact on the probability to find a donor. Roughly speaking if the size is double this probability may increase of 10% approximatively.

2. The impact of the selection mechanism is also very low. In the model 1 the efficiency difference between N_0 selection and optimal (implementable) selection is only 2 % of the size of the registry is multiplied by 2.

3. The optimal selection rule eliminates a very few number of very frequents types present in the initial registry but essentially eliminates numerous rare types.

5 Filter and implementable improvement of a registry

In the last section we have determined the optimal registry design but we have also shown that this optimal design is not implementable. We now consider implementable procedure based on presignaling or filtering which may be used in order to improve the registry. As in the previous section we start with an initial registry of N_0 donors and N_{oj} is the number of donors of type j. The total budget of the institution managing the registry is B and the cost of typing and of introducing a new donor in the registry is b.

A test is define by the following elements:

- it's cost c, typically smaller than b and the cost $b_1(\leq b)$ of introducing in the registry somebody who has been tested.
- a list of possible results of the test $\{1, ..., S\}$ and s is a possible result of the test.
- a joint distribution which represents the frequency in the population of potential donors of the test results and of the type j. This distribution is described by the probabilities

$$d(s, j) \quad s \in \{1, ..., S\} \quad j \in \{1, ..., J\}$$
$$d(s, j) \geq 0 \quad \sum_{s,j} d(s, j) = 1$$

This probability is assumed given. If the populations of donors and of patients are identical then the marginal distribution on j deduced from d, i.e. $d(s, j) = \sum_{s=1}^{S} d(s, j)$ should be equal to p_j.

In case of HLA typing we have in mind two example of tests. The first one is the observation of microsatellites on the two chromosomes. Micro satellites are characterised by loci between the genes on the 6th chromosome, more easy to observed then the HLA genes but strongly correlated to the HLA types. The information contained by microsatellites is "random" in the sense that the association between s and j is not perfect. In particular the conditional distribution $d(j|s)$ has in general a support not reduced to a singleton.

The other example we have in mind is the observation of SNP. Remember that a gene is characterised by a sequence of nucleotid (A,G,C or T) and it is possible to observe a subsequence only for each gene of the HLA system. The information contained by a given sequence of SNP is the exact knowledge of a partition of the list of types. In that case the conditional distribution of s given j has a support reduced to a singleton. However this is not true for the conditional distribution of j given s (which is the conditional distribution of the types given a subset). We will illustrate by examples below this two situations which appear to be particular cases of our general model.

Moreover let us remark that in this two cases, testing procedure requires the DNA extraction which is also the first step of typing. Then the testing step reduces the cost of typing and this explain the change of b into b_1.

A test procedure rises two questions : First how to use in the more efficient way the informations provided by the test ? Second is it efficient to implement the test in comparison with the random arrival of donors? More generally if several tests are available (no test is a particular case of these different choices) what is the best test in order to improve the registry?

Consider first the optimal use of the test information. A strategy is defined by three elements :

- A number N_1 of individuals who are directly introduced in the registry (i.e. typed without test)
- A number of individuals M who are tested.
- A sequence of probability $\sigma(s) \in [0, 1]$ for each possible value of the test result. The number $\sigma(s)$ represents the proportion or the probability of an individual with test result equal to s to be introduced in the registry and fully typed.

As the test as a cost, it is not obvious that all the potential donors should be tested before typing. It is in general optimal (if $b - b_1 < c$) to directly type a group of people and to use the test strategy in order to correct the natural arrival process of the different types. An advantage of this presentation is that the no test case is contained in our presentation ($M = 0$).

This specification contains pure strategies ($\sigma(s) = 0$ or 1) for which a potential donor is typed or not depending on the result of the test. In other way, a pure strategy is equivalent to a partition of the set $\{1, ..., S\}$ in two subsets $\mathcal{S}_0$ and $\mathcal{S}_1$. In that case if $s \in \mathcal{S}_0$ the potential donor is fully typed and introduced in the registry and if $s \in \mathcal{S}_1$ the process stops. As usual in decision theory it is powerful to consider mixed or random strategies where $\sigma(s)$ may be any element in $[0, 1]$. If s is observed the registry manager drawn between "typing" or "stop" with probabilities $\sigma(s)$ and $1 - \sigma(s)$.

Let us consider an initial registry ($N_{0j_{j=1,...,J}}$) and a test strategy defined by N_1, M and the $\sigma(s)$. This strategy defines:

- A random size of the increment of the registry:

$$N = N_1 + \left(\sum_{s=1}^{S} \sigma(s) d(s) \right) M$$

where $d(s) = \sum_{j=1}^{J} d(s,j)$ is the marginal probability of type s and $\sum_{s=1}^{S} \sigma(s)d(s)$ is the probability to be typed after a test.

- the frequency of type j deduced from the test strategy is equal to:

$$q_j = p_j \frac{N_1}{N} + \frac{\sum_{s=1}^{S} \sigma(s)d(s,j)}{\sum_{i=1}^{S} \sigma(s)d(s)} \left(\frac{N - N_1}{N} \right)$$

The first element of this sum corresponds to individuals introduced in the registry without pretest and the second correspond to the tested people.

The expected cost of a test strategy is equal to

$$bN_1 + cM + b_1 \left(\sum_{s=1}^{S} \sigma(s)d(s) \right) M$$

An optimal use of the information contain if a test is then obtained by minimizing

$$\sum_{j=1}^{J} p_j (1-a)^{N_{0j}} e^{-a\{p_j N_1 + \left(\sum_{s=1}^{S} \sigma(s)d(s,j)\right) M\}}$$

with respect to N_1, M and the $\sigma(s)$'s under the constraints

$$b_1 n N_1 + cM + b_1 \left(\sum_{s=1}^{S} \sigma(s)d(s) \right) M = B$$

$$N_1 \geq 0 \quad M \geq 0$$

$$\forall s = 1, ..., S \quad 0 \leq \sigma(s) \leq 1.$$

This optimization problem has no solution in closer form and should be done numerically. This numerical problem is however almost impossible to solve in practice due the dimension of S. An other important question is the knowledge of the joint distribution d. In the case of a "random" test like micro satellites informations, d should be derived from an estimation procedure based on a sample of individuals for whom s and j are observed. Here also the dimension of J and S is so large that no possible sample may provide a sufficient information about the joint distribution.

6 A Monte Carlo evaluation of an implementable improvement of a registry

As we have remarked in the previous section, the derivation of an optimal strategy based on a test in order to improve a registry faces to the curse of dimensionality. In the example of bone marrow transplant, the number of HLA types is extremely large (more then 60 000 types have been observed in France and don't constitute an exhaustive list) and the number of possible signals (observation of several microsatellites) is certainly greater than the number of unhabitants of France. The list of possible values of j and s are unknown and the joint probability $d(0, j)$ is very often calibrate using very small samples respectively to the number of possible values of the couple (s, j).

A realistic situation (corresponding to actual improvement problem of the french bone marrow registry) may be descried by the following arguments.

- we have a current registry, typically large from which a list of types and an evaluation of their probabilities may be derived. We assume here that donors and receivers are drawn from the same population.
- In this registry only type is available and not the signal s. However a sample of individuals is drawn (usually from the registy) for whom the signal is observable.

  In the bone marrow application, the current registry contains more than 110 000 individuals and the sample has a size below 5 000 (46...).

In that case the use of the sample to estimate the joint distribution $d(s, j)$ is impossible. However partial statistical analysis may be done. For example the HLA type is defined by three loci and the signal by 15 microsatellites. Partial analysis of dependence between are gene on one loci and a small (one to three) number of microsatellite is possible. We suggest the following procedure:

Step 1. The sample is usually not randomly generated from the registry, in particular in order to obtain some "rare" types. However it is necessary to construct a system of weights of individuals in the sample in order to have in the sample the same shape of the p_j as in reality.

Step 2: From partial statistical analysis we may derived a rule of thumb decision rule based on x, in our example, on the information given by microsatellite. If we restrict our attention to pure decision rule we should specify a function $\sigma(s)$ reduce in $\{0, 1\}$ for any sequence s of a list of micro satellite.

Step 3: Given this decision rule, we simulate an increment of the initial registry random arrival of *** are generated and only the one for which $\sigma(1) = 1$ are kept in the file up to a given size of the increment defined by budget constraint.

Step 4: A sample of patients are drawn for the population and the proportion available matching is compute.

Step 3 and 4 may simulated several times and provides a computation of the efficiency of the registry.

References

Hoffman-Smith C. (1993), Matching marrow donors & recipients: downsized system helps save more lives. Healthc Inform. Sep;10(9):18, 20.

Hurley CK, Schreuder GM, Marsh SG, Lau M, Middleton D, Noreen H. (1997), The search for HLA-matched donors: a summary of HLA-A,-B,-DRB1/3/4/5 alleles and their association with serologically defined HLA-A,-B,-DR antigens. Tissue Antigens. Oct;50(4):401-18.

Lonjou C, Clayton J, Cambon-Thomsen A, Raffoux C. (1995),HLA-A,-B,-DR haplotype frequencies in France implications for recruitment of potential bone marrow donors. Transplantation. Aug 27;60(4):375-83.

Muller CR. (2002) Computer applications in the search for unrelated stem cell donors.TransplImmunol.Aug;10(2-3):227-40.

Ottinger H, Grosse-Wilde M, Schmitz A, Grosse-Wilde H. (1994),Immunogenetic marrow donor search for 1012 patients: a retrospective analysis of strategies, outcome and costs. Bone Marrow Transplant. 1994-14 Suppl 4:S34-8.

Oudshoorn M, Cornelissen JJ, Fibbe WE, de Graeff-Meeder ER, Lie JL, Schreuder GM, Sintnicolaas K, Willemze R, Vossen JM, van Rood JJ. (1997), Problems and possible solutions in finding an unrelated bone marrow donor. Results of consecutive searches for 240 Dutch patients. Bone Marrow Transplant. Dec;20(12):1011-7.

Rendine S, Barbanti M, Borelli I, Dall'Omo AM, Roggero S, Sacchi N, Curtoni ES. (1999), The Italian Registry of Bone Marrow Donors: genetic structure and recruitment strategy - Ann Ist Super Sanita.35(1):21-34. Italian.

Schipper RF, Oudshoorn M, D'Amaro J, van der Zanden HG, de Lange P, Bakker JT, Bakker J, van Rood JJ. (1996), Validation of large data sets, an essential prerequisite for data analysis: an analytical survey of the Bone Marrow Donors Worldwide. Tissue Antigens. Mar;47(3):169-78.

Schuler U, Rutt C, Baier D, Keller JV, Stahr A, Grathwohl A, Ehninger G.(2000), Approaches to managing volunteer marrow donor registry HLA data. Algorithms for directing donor center-initiated HLA-DR typing of selected donors. Rev Immunogenet. 2000-2(4):541-6.

Speiser DE, Tiercy JM, Rufer N, Chapuis B, Morell A, Kern M, Gmur J, Gratwohl A, Roosnek E, Jeannet M.(1994),Relation between the resolution of HLA-typing and the chance of finding an unrelated bone marrow donor. Bone Marrow Transplant. 1994 Jun;13(6):805-9.

Sonnenberg FA, Eckman MH, Pauker SG. (1989),Bone marrow donor registries: the relation between registry size and probability of finding complete and partial matches. Blood. 1989 Nov 15;74(7):2569-78.

Takahashi K, Juji T, Miyazaki H. (1989), Determination of an appropriate size of unrelated donor pool to be registered for HLA-matched bone marrow transplantation. Transfusion. 1989 May;29(4):311-6.

Tiercy JM, Bujan-Lose M, Chapuis B, Gratwohl A, Gmur J, Seger R, Kern M, Morell A, Roosnek E. (2000), Bone marrow transplantation with unrelated donors: what is the probability of identifying an HLA-A/B/Cw/DRB1/B3/B5/DQB1-matched donor? Bone Marrow Transplant. 2000 Aug;26(4):437-41.

Tiercy JM, Stadelmann S, Chapuis B, Gratwohl A, Schanz U, Seger RA, Faveri GN, Kern M, Morell A, Schwabe R, Schneider P. (2003), Quality control of a national bone marrow donor registry: results of a pilot study and proposal for a standardized approach. Bone Marrow Transplant. 2003 Sep;32(6):623-7.