Probability in the Engineering and Informational Sciences, page 1 of 21, 2018.
doi:10.1017/S0269964818000153

BLOOD GROUP PROBABILITIES BY NEXT OF KIN

Joost H. J. van Sambeeck
Department of Transfusion Technology Assessment, Sanquin, Amsterdam, The Netherlands and Centre for Healthcare Operations Improvement & Research, University of Twente, Enschede, The Netherlands
E-mail: j.vansambeeck@sanquin.nl

Nico M. van Dijk
Department of Stochastic Operations Research, University of Twente, Enschede, The Netherlands and Centre for Healthcare Operations Improvement & Research, University of Twente, Enschede, The Netherlands
E-mail: n.m.vandijk@utwente.nl

Wim L. A. M. de Kort
Department of Donor Studies, Sanquin, Amsterdam, The Netherlands
and
Department of Social Medicine, Academic Medical Center, Amsterdam, The Netherlands
E-mail: w.dekort@sanquin.nl

Henk Schonewille
Department of Experimental Immunohematology, Sanquin, Amsterdam, The Netherlands
E-mail: h.schonewille@sanquin.nl

Mart P. Janssen
Department of Transfusion Technology Assessment, Sanquin, Amsterdam, The Netherlands and Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
E-mail: m.janssen@sanquin.nl

For rare blood groups the recruitment of donor relatives, for example siblings, is expected to be effective, since the probability of a similar rare blood group is likely. However, the likelihood differs between blood groups and is not commonly available. This paper provides a unified mathematical formulation to calculate such likelihoods. From a mathematical and probabilistic point of view, it is shown that these likelihoods can be obtained from the computation of a stationary genotype distribution. This, in turn, can be brought down to a system of quadratic stochastic operators. A generic mathematical approach is presented which directly leads to a stationary genotype distribution for arbitrary blood groups. The approach enables an exact computation for the effectiveness of recruiting next of kin for blood donorship. Next to an illustration of computations for 'standard' ABO and Rhesus-D blood groups, it is particularly illustrated for the extended Rhesus blood group system. Also other applications requiring next of kin blood group associations can be solved directly by using the unified mathematical formulation.

Keywords: applied probability, operations research, stochastic modeling

1. INTRODUCTION

1.1. Motivation

The challenges faced by blood transfusion services are becoming more complex and are changing continuously over time due to growing economic pressure, new technologies, and increasing customer expectations [5,16]. One of these expectations is the ability to select extensively (blood group) matched red blood cells (RBCs) for transfusion recipients, to decrease the number and severity of transfusion reactions. However, current blood donor recruitment strategies are based on historical matching strategies and cannot meet the demand for extensively matched blood products. Furthermore, due to increasing immigration rates and differences in blood group distributions between ethnic populations the diversity among blood groups within the transfusion population increases. For instance, the blood group profiles of Caucasian individuals (i.e., individuals with European ancestors) and individuals from African descent differ significantly. In contrast, in the donor base composition ethnic minorities are underrepresented, complicating extended blood group matching of donors and transfusion recipients. Hence, one of the major challenges for current blood donor recruitment practice is to maintain an adequate donor base with a sufficiently diverse blood group composition [3]. In actual fact, an overrepresentation of donors from African descent would be preferable, as individuals from African descent have a higher probability of requiring repeated blood transfusions as a result of sickle cell decease, which is uncommon in other populations [2].

In practice, it has been shown effective to increase the number of donors with O, Rhesus-D (RhD) negative blood groups by recruiting among their relatives, since these are more likely to be O, RhD-negative than individuals in the general population. Although intuitively this seems to be an effective strategy, it is not evident to what extent such strategies are more effective than random donor selection. Moreover, it gives rise to the question whether this also holds for other blood group combinations. If so, it may steer towards more effective recruitment strategies.

1.2. Approach

To model blood group antigen inheritance quadratic stochastic operators (QSOs) are used, as introduced by Bernstein in 1924 [1]. Recently, Ganikhodjaev et al. [7–9] applied QSOs to model the heredity of ABO and RhD blood groups. However, a general formulation that goes beyond the standard ABO, RhD blood groups was not given. In addition, an exact computation of the effectiveness of recruiting relatives of donors with rare blood groups has not been included. Of course, the idea that relatives have similar blood groups is intuitively correct, but quantification is insightful and allows balancing recruitment efforts against the benefits from blood group matching.

In this paper, therefore, we present a unified mathematical formulation to determine the probability that two relatives (next of kins) share the same blood group. In short, the steps and formulation that will be provided, transform phenotype distributions into genotype distributions and back. By this generic mathematical approach we can directly analyze the effectiveness of specific next of kin recruitment strategies, for any blood group, ethnicity, and population (as numbers may differ worldwide).

The mathematical approach only requires a phenotype distribution as an input, whereas the population genotype distribution is required for calculating the blood group distribution probability for the next of kin. Phenotype distributions can be easily determined by simple blood tests, genotype distributions are more difficult to obtain. However, these genotype distribution can be derived from the phenotype distributions using our generic mathematical approach.

This paper starts with a known, but motivational example in Section 2. Next, in Section 3, a unified mathematical formulation of the approach is covered. In Section 4, this unified mathematical formulation is used to compute the effectiveness of recruiting next of kin for blood donorship. Finally, we explore some specific applications of the approach in Section 5. At the end of the paper, we provide a clear overview of the notation used (see Appendix A).

2. MOTIVATIONAL AND ILLUSTRATIVE EXAMPLE

In this section, let us first provide the genetic terminology and illustrate our steps and formulation for the ‘standard’ ABO and RhD blood groups. That is, we show

- how the approach for determining the distribution of genotypes in a population essentially comes down to a system of quadratic equations,
- how the distribution of genotypes can be used to evaluate the effectiveness of targeted recruitment strategies for the ABO and RhD blood groups separately,
- how the results for both blood groups can be combined.

Later, in Sections 3 and 4, the same steps and approach are provided in a unified mathematical formulation, such that this formulation can be applied to any blood group system.

2.1. ABO, RhD blood groups

According to the International Society of Blood Transfusion (ISBT), there are more than 300 different blood group antigens belonging to 35 blood group systems [12]. Each antigen can be either present or absent on the surface of an RBC, leading to an extremely large number of different blood group profiles. In practice, however, not all antigens are equally important with regard to transfusion-related problems. The most important antigens are A and B (both belonging to the ABO blood group system), followed by RhD, which belongs to the RhD blood group system. Taking only these three antigens into consideration the total number of blood group profiles can be compressed into eight major groups, the so-called ABO, RhD blood groups. These ABO, RhD blood groups consist of a combination of a blood group belonging to the ABO blood group system (O, A, B, AB) and a RhD blood group (RhD-neg (d), RhD-pos (D)).

For just the RhD blood groups three different genotypes ($\mathcal{G}_{\mathrm{D}} = \{dd, Dd, DD\}$) and two different phenotypes ($\mathcal{F}_{\mathrm{D}} = \{\mathrm{d}, \mathrm{D}\}$) exist, where the genotype is a genetic code that determines which antigen might be expressed on the surface of the RBCs. The expression of particular antigen is called the phenotype. Moreover, multiple genotypes may lead to the same phenotype. The relation between the different RhD genotypes and phenotypes is shown in the following matrix:

$$S = \begin{array}{c} \\ dd \\ Dd \\ DD \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{d} & \mathrm{D} \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \end{array}, \tag{2.1}$$

where 1 indicates which genotypes results in a particular phenotype. Note that genotypes (and genes) are presented in italics and phenotypes (and antigens) are presented in a regular typeface.

TABLE 1. Phenotype frequencies for the ABO–RhD blood groups [13].

		O	A	B	AB
		0.48	0.40	0.09	0.03
d	0.15	0.072	0.060	0.013	0.005
D	0.85	0.408	0.340	0.077	0.026

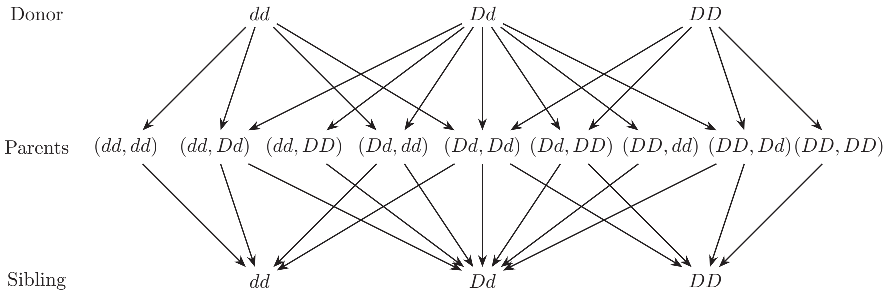


FIGURE 1. Probability diagram which relates the RhD genotype of a donor to the RhD genotype of its parents and siblings.

Similarly, the ABO blood group system consists of six different genotypes ($\mathcal{G}_{\mathrm{ABO}} = \{OO, OA, OB, AA, AB, BB\}$) and four different phenotypes ($\mathcal{F}_{\mathrm{ABO}} = \{\mathrm{O}, \mathrm{A}, \mathrm{B}, \mathrm{AB}\}$). The phenotype frequencies for the ABO, RhD blood groups in the general Caucasian population are given in Table 1. The RhD and ABO blood groups belong to two blood group systems and are inherited independently. Therefore, in the next sections we will explore which steps are required to investigate the effectiveness of recruiting next of kin with respect to the RhD and ABO blood groups separately. At the end of this section the results for both blood groups are combined.

Note that most of the computations performed in this section are similar to what can be found in the literature [4,7–10,14]. However, the specific structure of the mathematical approach, the usage of just a known phenotype distribution, and the connection to the effectiveness of targeted recruitment strategies (see Section 2.4) are new.

2.2. Motivational example for the RhD blood group

Figure 1 shows a probability diagram describing the relation between the RhD genotype of a donor and its parents and siblings (i.e. brothers or sisters). The probability that a donor has a particular genotype is the a priori probability. From the figure it is clear that the probability of a sibling having the same genotype requires information on genotypes of the parents. However, it might be that the distribution of genotypes in the general population is unknown or difficult to obtain. On the other hand, the phenotype distribution for the general population is usually more easily available, so it would be convenient if we could use this instead, to determine the genotype distribution. This is possible by using quadratic stochastic operators.

When the a priori probabilities are known, Bayes rule is applied to find the probability that a relative of a donor has a specific RhD genotype, given the genotype of the donor. In order to compute these probabilities, particularly for a sibling of a donor, we thus need to work top-down.

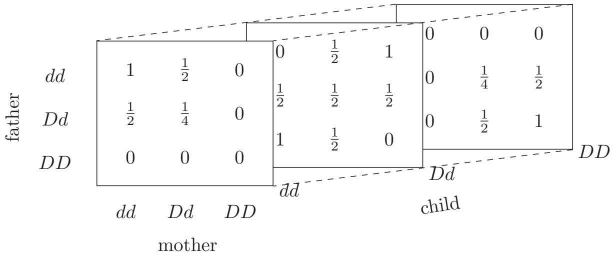


FIGURE 2. Inheritance matrix P for the RhD blood group.

2.3. Finding a stationary distribution

For a particular blood group, a child inherits its genotype from a combination of genotypes of the parents. For the RhD blood group a genotype consists of two genes, each of which either d or D, leading to three possible genotype combinations: dd, Dd, and DD. Each parent gives one of these two to the child. The probability that two parents with a particular genotype conceive a child with a certain genotype is captured by an inheritance matrix P. For the RhD blood groups, the inheritance matrix is depicted in Figure 2. We are going to use this inheritance matrix $P \in \mathbb{R}^{3\times3\times3}$ to compute a stationary distribution of genotypes. The exact structure of this matrix will be explained in Section 3.1.

Let $\boldsymbol{x}^{(n)} \in \mathbb{R}^{3\times1}$ be a column vector containing the genotype distribution for the RhD blood groups in generation n. We assume that this genotype distribution is stationary, which implies that the distribution of genotypes in generation $n-1$ is equal to the distribution of genotypes in generation n: $\boldsymbol{x}^{(n-1)} = \boldsymbol{x}^{(n)} = \boldsymbol{x}$. Let $\boldsymbol{x}_{\text{father}}$, $\boldsymbol{x}_{\text{mother}}$, and $\boldsymbol{x}_{\text{child}}$ be the genotype distributions of respectively father, mother, and child. Then, in a stationary population, the following equations hold:

$$\boldsymbol{x}_{\text{father}}^\top P \boldsymbol{x}_{\text{mother}} = \boldsymbol{x}_{child} \quad \Rightarrow \quad \boldsymbol{x}^\top P \boldsymbol{x} = \boldsymbol{x}. \tag{2.2}$$

Moreover, the genotypes are related to the phenotypes. This relation was given in the matrix S (see Eq. (2.1)). Besides Eq. (2.2) the following equation should also hold for variable $\boldsymbol{x}$:

$$S^\top \boldsymbol{x} = \boldsymbol{f}, \tag{2.3}$$

where $\boldsymbol{f} \in \mathbb{R}^{2\times1}$ is the phenotype distribution. For the RhD blood groups, Eqs. (2.2) and (2.3) can be solved analytically, which gives:

$$\begin{cases} x_{dd}^2 + x_{dd}x_{Dd} + \frac{1}{4}x_{Dd}^2 & = \quad x_{dd} \\ x_{dd}x_{Dd} + 2x_{dd}x_{DD} + \frac{1}{2}x_{Dd}^2 + x_{Dd}x_{DD} & = \quad x_{Dd} \\ \frac{1}{4}x_{Dd}^2 + x_{Dd}x_{DD} + x_{DD}^2 & = \quad x_{DD} \\ x_{dd} & = \quad f_d \\ x_{Dd} + x_{DD} & = \quad f_D \end{cases} \Rightarrow \begin{cases} x_{dd} & = \quad f_d \\ x_{Dd} & = \quad f_D - \left(1 - \sqrt{f_d}\right)^2 \\ x_{DD} & = \quad \left(1 - \sqrt{f_d}\right)^2 \end{cases}.$$

Note that this analytic solution is in accordance with the Hardy–Weinberg law [11]. Since $\boldsymbol{f}^\top = (f_d, f_D) = (0.15, 0.85)$ we get

$$\boldsymbol{x} = \begin{bmatrix} x_{dd} \\ x_{Dd} \\ x_{DD} \end{bmatrix} = \begin{bmatrix} 0.150 \\ 0.475 \\ 0.375 \end{bmatrix}.$$

In a similar way, equations and computations can be provided for the ABO-blood group system from which we find

$$\boldsymbol{x} = \begin{bmatrix} x_{OO} \\ x_{OA} \\ x_{OB} \\ x_{AA} \\ x_{AB} \\ x_{BB} \end{bmatrix} = \begin{bmatrix} 0.480 \\ 0.341 \\ 0.084 \\ 0.061 \\ 0.030 \\ 0.004 \end{bmatrix}.$$

In Casas et al. [4] square root expressions have been provided for the ABO blood group system and are therefore omitted here. However, this reference has not discussed the concept of effectiveness. This will be elaborated on in the next section.

2.4. Effectiveness of recruiting next of kin for donorship

Donors are recruited for their phenotypes expressions (blood is matched on phenotypes), however, inheritance is determined by genotypes. Therefore, to compute the probability that a sibling of a donor with a particular phenotype has the same phenotype, the stationary genotype distribution is required. Once this genotype distribution has been obtained the likelihood of a particular blood group for a sibling, given the blood group of a relative, can then be computed using Bayes' rule.

Suppose that we have a RhD-pos donor, the likelihood that its sibling is also RhD-pos can be computed by calculating the following conditional probability:

$$\mathbb{P}\left[\text{sibling D} \mid \text{donor D}\right] = 0.908.$$

Details on this calculation are provided in Appendix B.

We find that the conditional probability is slightly higher than the probability that a random individual is RhD-pos (0.850). The effectiveness, defined as the difference between these two probabilities, is equal to

$$\begin{aligned} E_D &= \mathbb{P}\left[\text{sibling D} \mid \text{donor D}\right] - f_D \\ &= 0.058. \end{aligned}$$

Figure 3 shows the results of an analysis of ABO and RhD blood groups. Especially for rare blood groups (i.e. B, AB, and RhD-neg) it appears to be effective to recruit among relatives. Here, the likelihood of a similar blood group is considerably higher than that of the general population. For example, for the RhD-neg blood group the likelihood increases from 0.16 to 0.40 for parents and to 0.48 for siblings. Note that the probability of the siblings is higher than that of the parents.

The most important ABO–RhD blood group is O, RhD-neg, since this is the blood group of a so-called universal donor. This means that every individual can receive RBCs from a donor with this blood group. Figure 3 show that recruiting O, RhD-neg donors among relatives of donors with an O, RhD-neg blood group is five or four times more effective for siblings and parents respectively, than recruiting donors at random. These computations are insightful when assessing targeted donor recruitment among relatives.

This section provided an illustration of calculating next of kin blood group probabilities for the ABO–RhD blood groups. In the next sections we will provide a more generic mathematical framework to compute (i) stationary genotype distribution and (ii) effectiveness of recruiting next of kin for blood donorship by using QSOs and Bayesian statistics. This allows calculating next of kin probabilities for more complex blood group combinations.

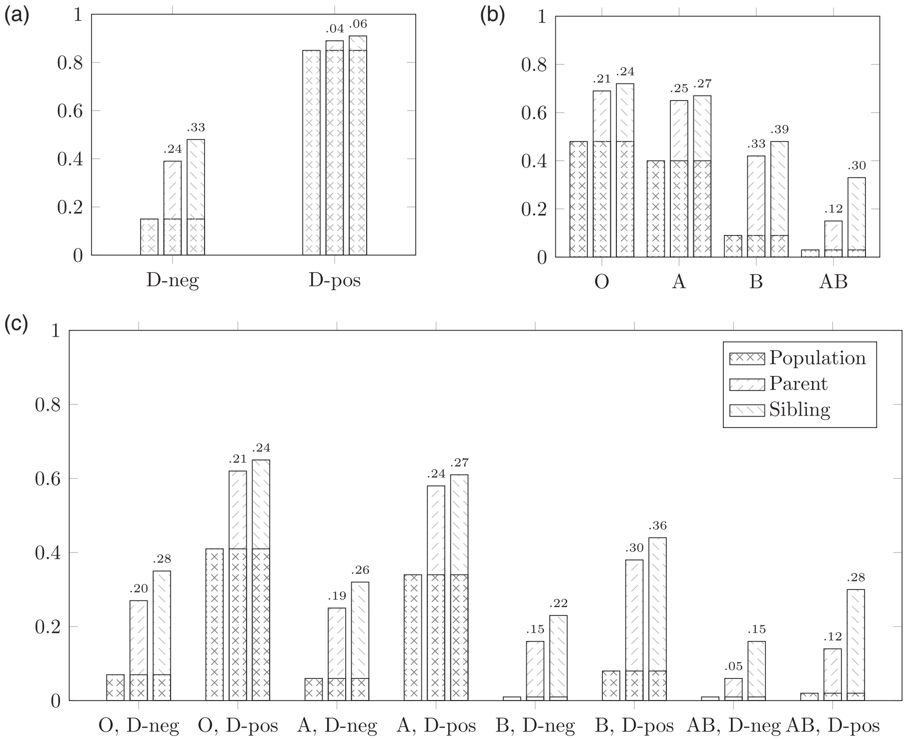


FIGURE **3.** Proportion of ABO–RhD blood groups in the general population and conditional probabilities that parents/siblings have the same blood group. The numbers above the conditional probabilities represent the effectiveness of recruiting relatives of a donor with a known blood group, where the effectiveness is defined as the difference between the the proportion of individuals with a particular blood group in the population (see Table 1) and the conditional probabilities. (a) Proportion and inheritance probability of RhD blood groups. (b) Proportion and inheritance probability of ABO blood groups. (c) Proportion and inheritance probability of ABO–RhD blood groups.

3. GENERIC MATHEMATICAL APPROACH

As illustrated in Section 2.4 we want to calculate the conditional probability that a relative of a donor has the same phenotype as the donor, or mathematically stated:

$$\mathbb{P}\left[\text{relative } \varphi \mid \text{donor } \varphi\right], \tag{3.1}$$

where $\varphi \in \mathcal{F}$ is the known phenotype of the donor. We, therefore, aim to provide a unified mathematical framework, starting in this section with providing a generic mathematical approach for computing the stationary genotype distribution. In the next section, it will be shown how this stationary genotype distribution is used to calculate the effectiveness of recruiting next of kin for blood donorship.

First, in Section 3.1, we start with mathematically modeling the blood group genetics and introduce some notation. Next, In Section 3.2, the calculation steps required are listed and in the remainder of Section 3, we elaborate on the computation of a stationary genotype distribution.

3.1. Blood group genetics

To explain the relation between the blood group of a child and its parents, we start with a compact description of the underlying genetic mechanism of inheritance, based upon the terminology of Elston et al. [6]. The information of an individual's blood group is present on the genes, which occur in pairs on homologous chromosomes at particular positions called loci (singular: locus). Genes that occur at the same locus are allelic to each other and are, therefore, also referred to as alleles. Each allele may encode for the production of a specific antigen. For example, the ABO blood groups are determined by three alleles A, B, and O, where A encodes for the production of antigen A, B encodes for the production of antigen B, and O encodes for no antigen production. To write this down mathematically, we first introduce for each locus a set of alleles $\mathcal{L}$ and a set of antigens $\mathcal{A}$. Then, for each allele $l \in \mathcal{L}$ a binary vector of length $|\mathcal{A}|$ is constructed with $l_a = 1$ if allele l encodes for the production of antigen $a \in \mathcal{A}$ and $l_a = 0$ otherwise. Finally, the alleles are sorted into a lexicographical order, for example $\mathcal{L}_{\mathrm{ABO}} = \{A, B, O\} = \{\{1, 0\}, \{0, 1\}, \{0, 0\}\}$. Hence, $\mathcal{L}$ is a lexicographical ordered set of alleles.

In contrast to the ABO blood groups, which are determined by alleles lying on a single locus, the Rh blood groups are determined by a combination of alleles occurring at multiple loci. This combination of alleles is called a haplotype (the multilocus analogue of an allele at a single locus), where a haplotype consists of one allele from each of the loci. The set of haplotypes is denoted by $\mathcal{H}$, where each $h \in \mathcal{H}$ can be written as a union of alleles belonging to unique loci. For example, the set of haplotypes for the Rh blood groups is determined by alleles from three loci ($\mathcal{L}_{\mathrm{C}} = \{C, c\} = \{\{1, 0\}, \{0, 1\}\}$, $\mathcal{L}_{\mathrm{D}} = \{D, d\} = \{\{1\}, \{0\}\}$, and $\mathcal{L}_{\mathrm{E}} = \{E, e\} = \{\{1, 0\}, \{0, 1\}\}$) leading to eight different Rh haplotypes:

Binary representation	Antigens	Haplotype
$\{1,0,1,1,0\}$	CDE	CDE
$\{1,0,1,0,1\}$	CDe	CDe
$\{0,1,1,1,0\}$	cDE	cDE
$\{0,1,1,0,1\}$	cDe	cDe
$\{1,0,0,1,0\}$	CE	CdE
$\{1,0,0,0,1\}$	Ce	Cde
$\{0,1,0,1,0\}$	cE	cdE
$\{0,1,0,0,1\}$	ce	cde

Although the sets $\mathcal{L}_{\mathrm{C}}$, $\mathcal{L}_{\mathrm{D}}$, and $\mathcal{L}_{\mathrm{E}}$ all consist of two alleles, they are different. On the one hand, the alleles in the sets $\mathcal{L}_{\mathrm{C}}$ and $\mathcal{L}_{\mathrm{E}}$ always lead to the production of antigens, that is $\{1, 0\} \in \mathcal{L}_{\mathrm{C}}$ implies production of antigens C, $\{0, 1\} \in \mathcal{L}_{\mathrm{C}}$ implies production of antigens c, $\{1, 0\} \in \mathcal{L}_{\mathrm{E}}$ implies production of antigens E, and $\{0, 1\} \in \mathcal{L}_{\mathrm{E}}$ implies production of antigens e. On the other hand, the alleles in the set $\mathcal{L}_{\mathrm{D}}$ might lead to the production of an antigen, that is $\{1\} \in \mathcal{L}_{\mathrm{D}}$ implies production of antigens D, but $\{0\} \in \mathcal{L}_{\mathrm{D}}$ implies that no antigens are produced. We define $\mathcal{H}$ as a lexicographical ordered set of haplotypes with cardinality $|\mathcal{H}| = \prod_i |\mathcal{L}_i|$.

Let $\mathcal{G} = \{\gamma_1, \ldots, \gamma_m\}$ be the lexicographic ordered set of genotypes consisting of all combinations of 2 haplotypes from $\mathcal{H}$:

$$\mathcal{G} = \{ \underbrace{\{h_1, h_1\}, \ldots, \{h_1, h_{|\mathcal{H}|}\}}_{|\mathcal{H}| \text{ elements}}, \underbrace{\{h_2, h_2\}, \ldots, \{h_2, h_{|\mathcal{H}|}\}}_{|\mathcal{H}|-1 \text{ elements}}, \ldots, \underbrace{\{h_{|\mathcal{H}|}, h_{|\mathcal{H}|}\}}_{1 \text{ element}}\}, \tag{3.2}$$

with cardinality $m = |\mathcal{G}| = (1/2)|\mathcal{H}|(|\mathcal{H}| + 1)$. Finally, let $\mathcal{F} = \{\varphi_1, \ldots, \varphi_n\}$ be the lexicographic ordered set of phenotypes, with cardinality $n = |\mathcal{F}|$. These phenotypes determine which antigens are present on the RBCs. Let $S \in \{0, 1\}^{m\times n}$ be a matrix describing the relation between genotypes and phenotypes, that is

$$S_{ij} = \begin{cases} 1, & \text{if genotype } \gamma_i \in \mathcal{G} \text{ leads to phenotype } \varphi_j \in \mathcal{F}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

Children inherit blood group antigens from their parents. Which antigens are inherited depends on the genotypes of both parents. Suppose that the father has genotype $\gamma_i \in \mathcal{G}$ ($\gamma_i = \{h_{i_1}, h_{i_2}\}$), the mother has genotype $\gamma_j \in \mathcal{G}$ ($\gamma_j = \{h_{j_1}, h_{j_2}\}$), and they get a child with genotype $\gamma_k \in \mathcal{G}$. Clearly, this child could have four different genotypes, since there are four different combinations of the haplotypes of the parents: $\{h_{i_1}, h_{j_1}\}$, $\{h_{i_1}, h_{j_2}\}$, $\{h_{i_2}, h_{j_1}\}$, and $\{h_{i_2}, h_{j_2}\}$. Without loss of generality, we assume that Mendelian rules hold, which implies that each combination occurs with probability 1/4. In this section, we will index the genotypes by

- γ_i - genotype of the father,
- γ_j - genotype of the mother,
- γ_k - genotype of the child,

and use no index if we do not refer specifically to a father, mother, or child.

In order to construct the inheritance matrix $P \in \mathbb{R}^{m\times m\times m}$, we first introduce some vectors $v_h = [v_h(\gamma_1), \ldots, v_h(\gamma_m)]$, $h \in \mathcal{H}$, where $v_h(\gamma_i)$ is the probability that a parent with genotype $\gamma_i = \{h_{i_1}, h_{i_2}\}$ will give haplotype h to the child:

$$v_h(\gamma_i) = \begin{cases} 1, & \text{if } \gamma_i = \{h, h\}, \\ \frac{1}{2}, & \text{if } \gamma_i = \{h, \not h\} \text{ or } \gamma_i = \{\not h, h\}, \\ 0, & \text{otherwise.} \end{cases} \qquad h \in \mathcal{H}, \quad \gamma_i \in \mathcal{G}, \tag{3.4}$$

Then the probability that a child has genotype $\gamma_k = \{h_i, h_j\}$, where h_i is the haplotype the child inherited from the father and h_j is the haplotype the child inherited from the mother, is equal to:

$$P(\gamma_k) = P(\{h_i, h_j\}) = \begin{cases} \boldsymbol{v}_{h_i}\boldsymbol{v}_{h_j}^T, & \text{if } i = j, \\ \boldsymbol{v}_{h_i}\boldsymbol{v}_{h_j}^T + \boldsymbol{v}_{h_j}\boldsymbol{v}_{h_i}^T, & \text{if } i \neq j. \end{cases} \tag{3.5}$$

Note that $P(\gamma_k) \in \mathbb{R}^{m\times m}$ is a two-dimensional matrix as is shown in Figure 4.

3.2. Steps

The probability that two relatives share the same blood group is substantially higher than the probability that two individuals from the general population share the same blood group. For selective donor recruitment it is therefore worthwhile to quantify these probabilities as a function of the family relation. One might have the perception that these probabilities can be easily computed by elementary statistics. This is true, except that the a priori probabilities, that is the genotype distributions, are generally unknown and have to be calculated first. As will be shown, these a priori probabilities can be determined by a system of quadratic equations or rather a system of quadratic stochastic operators. Therefore, the mathematical

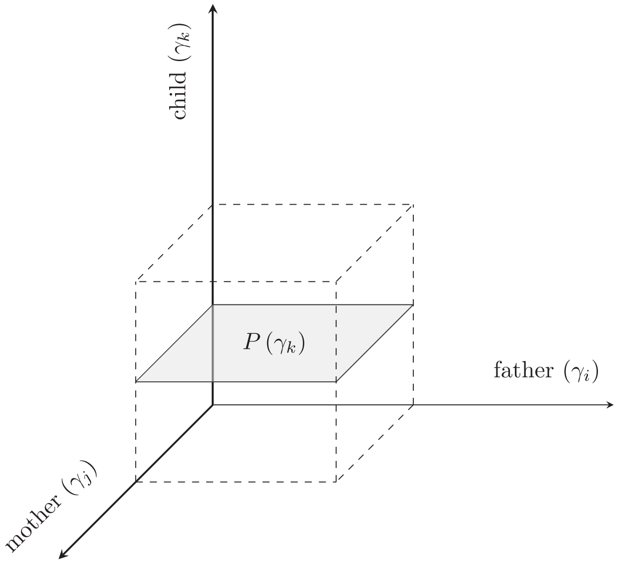


FIGURE 4. Inheritance matrix P, with $P(\gamma_k \mid \gamma_i, \gamma_j)$ the probability that two individuals with genotypes $\gamma_i \in \mathcal{G}$ and $\gamma_j \in \mathcal{G}$ conceive a child with genotype $\gamma_k \in \mathcal{G}$.

approach, combining both elementary statistics and operations research-related methods, can be divided into the following three steps:

- Determine the stationary distribution of genotypes.
- Compute the probability that a relative of a donor has a particular phenotype given that this donor has a particular phenotype.
- Compute the effectiveness of recruiting a next of kin donor instead of an individual from the general population.

3.3. Determine a stationary distribution of genotypes

By performing simple tests it is possible to determine the distribution of phenotypes in a population for a (combination of) blood group(s). Genotype distributions or allele frequencies are more difficult to obtain. A way to obtain estimates of these genotype distributions is by using QSOs. These estimates are based on the known phenotype distributions and the assumption that the genotype distributions within a population are stable.

First, we explain how we can model the inheritance of antigens by using a quadratic stochastic operator. This leads to a system of quadratic equations. Next, we show how this system of quadratic equations can be solved, by iteratively solving a least-squares problem.

Consider a set $\mathcal{G}$ of genotypes. Let x_γ be a variable that describes the frequency of genotype $\gamma \in \mathcal{G}$ in a population and let $P(\gamma_k \mid \gamma_i, \gamma_j)$ be the probability that two individuals with genotypes $\gamma_i \in \mathcal{G}$ and $\gamma_j \in \mathcal{G}$ conceive a child with genotype $\gamma_k \in \mathcal{G}$. Now, as in Section 2.3 using $\boldsymbol{x} = \boldsymbol{x}_{\text{father}} = \boldsymbol{x}_{\text{mother}} = \boldsymbol{x}_{\text{child}}$, the following equations hold:

$$\boldsymbol{x}_{\text{father}}^\top P \boldsymbol{x}_{\text{mother}} = \boldsymbol{x}_{\text{child}} \Rightarrow \boldsymbol{x}^\top P \boldsymbol{x} = \boldsymbol{x}, \tag{3.6}$$

where $P(\gamma_k \mid \gamma_i, \gamma_j)$ is the heredity matrix satisfying $P(\gamma_k \mid \gamma_i, \gamma_j) \geq 0$, $P(\gamma_k \mid \gamma_i, \gamma_j) = P(\gamma_k \mid \gamma_j, \gamma_i)$, $\sum_{\gamma_k \in \mathcal{G}} P(\gamma_k \mid \gamma_i, \gamma_j) = 1$.

Since we have a system of quadratic equations, there could be multiple stationary solutions $\boldsymbol{x}$. Based on the phenotype distribution $\boldsymbol{f}$, we can investigate which of these solutions is correct, requiring $S^\top \boldsymbol{x} = \boldsymbol{f}$. Hence, we need to solve the following system of equations:

$$\begin{cases} \boldsymbol{x}^\top P \boldsymbol{x} = \boldsymbol{x}, \\ S^\top \boldsymbol{x} = \boldsymbol{f}. \end{cases} \tag{3.7}$$

To compute a solution $\boldsymbol{x}$ that satisfies (3.7), we are first going to rewrite this system of quadratic equations as:

$$\begin{cases} \boldsymbol{x}^\top P \boldsymbol{x} = \boldsymbol{x} \\ S^\top \boldsymbol{x} = \boldsymbol{f} \end{cases} \quad \Rightarrow \quad \begin{cases} \left(\boldsymbol{x}^\top P - I\right) \boldsymbol{x} = \boldsymbol{0} \\ S^\top \boldsymbol{x} = \boldsymbol{f} \end{cases} \quad \Rightarrow \quad \underbrace{\begin{bmatrix} \boldsymbol{x}^\top P - I \\ S^\top \end{bmatrix}}_{A(\boldsymbol{x})} \boldsymbol{x} = \underbrace{\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{f} \end{bmatrix}}_{b},$$

where $I \in \mathbb{R}^{m \times m}$ is the identity matrix and $\boldsymbol{0} \in \mathbb{R}^{m \times 1}$ is the zero vector. In short, we thus get

$$A(\boldsymbol{x})\boldsymbol{x} = \boldsymbol{b}, \tag{3.8}$$

with $A(x) \in \mathbb{R}^{(m+n)\times m}$, $\boldsymbol{x} \in \mathbb{R}^{m \times 1}$, and $\boldsymbol{b} \in \mathbb{R}^{(m+n)\times 1}$. Since this is not a linear but an implicit equation, an iterative approximate procedure is proposed. For given $\boldsymbol{x}_0$ let $\boldsymbol{x}^{(n)}$ for $n = 1, 2, \ldots$ be determined by

$$A(\boldsymbol{x}^{(n-1)})\boldsymbol{x}^{(n)} = b. \tag{3.9}$$

In Section 3.5 we will make this more explicit. To this end, since there are different methods possible, let us first provide the one that will be used.

3.4. QR-factorization

If we would regard the matrix A independent of $\boldsymbol{x}^{(n-1)}$, then we just have a system of linear equations. Normally, a linear system can easily be solved by applying Gaussian elimination. However, in this specific case, an exact solution may not exist, since this system has more equations than unknown variables. Accordingly, we propose to solve a least-squares problem. Let $\boldsymbol{r} = A\boldsymbol{x}^{(n)} - \boldsymbol{b}$, or simply $\boldsymbol{r} = A\boldsymbol{x} - \boldsymbol{b}$, be the vector of residuals. Next, we want to find a solution $\boldsymbol{x}$ that minimizes the sum of squared residuals:

$$\min_{\boldsymbol{x} \in \mathbb{R}^m} \left\{ \|\boldsymbol{r}\|_2^2 \mid \boldsymbol{r} = A\boldsymbol{x} - \boldsymbol{b} \right\}. \tag{3.10}$$

Different methods are known to solve least-squares problems. One of them, based upon QR factorization [15], will be applied here. If the matrix A has full column rank, then it can be decomposed into the matrices Q and R ($A = QR$), such that the matrix $Q \in \mathbb{R}^{(m+n)\times(m+n)}$ has orthonormal columns and the matrix $R \in \mathbb{R}^{(m+n)\times m}$ is upper triangular (see Figure 5). In Appendix C, we proof that the matrix A has indeed full column rank and hence the residuals can be written as $\boldsymbol{r} = QR\boldsymbol{x} - \boldsymbol{b}$.

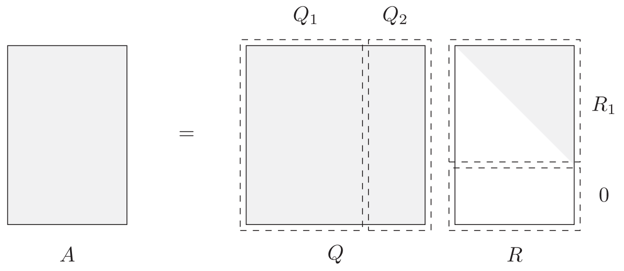


FIGURE 5. QR decomposition of the matrix A, where Q is a orthogonal matrix and R is an upper triangular matrix. Matrices Q and R can be partitioned, such that $Q = [Q_1 \quad Q_2]$ and $R^\top = [R_1 \quad 0]$.

Define $\bar{\boldsymbol{r}} = Q^\top \boldsymbol{r}$ as a linear transformation of the residuals. Then minimizing $\|\bar{\boldsymbol{r}}\|_2^2$ is equivalent to minimizing $\|\boldsymbol{r}\|_2^2$, since

$$\|\bar{\boldsymbol{r}}\|_2^2 = \left(Q^\top \boldsymbol{r}\right)^\top Q^\top \boldsymbol{r} = \boldsymbol{r}^\top Q Q^\top \boldsymbol{r} = \boldsymbol{r}^\top \boldsymbol{r} = \|\boldsymbol{r}\|_2^2 .$$

Moreover, note that $\bar{\boldsymbol{r}} = R\boldsymbol{x} - Q^\top \boldsymbol{b}$ and hence (3.10) is equivalent to

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \left\{ \|\bar{\boldsymbol{r}}\|_2^2 \;\middle|\; \bar{\boldsymbol{r}} = R\boldsymbol{x} - Q^\top \boldsymbol{b} \right\}. \tag{3.11}$$

We can find an exact solution to (3.11) by exploiting the specific structure of R. Partition R into an upper triangular matrix R_1 and a zero matrix. Similarly, we can write $Q = [Q_1 \quad Q_2]$ and $\bar{\boldsymbol{r}}^\top = [\bar{\boldsymbol{r}}_1^\top \quad \bar{\boldsymbol{r}}_2^\top]$ (see Figure 5). Hence, $\bar{\boldsymbol{r}} = R\boldsymbol{x} - Q^\top \boldsymbol{b}$ can be split into two sets of equations

$$\begin{cases} \bar{\boldsymbol{r}}_1 = R_1 \boldsymbol{x} - Q_1^\top \boldsymbol{b}, \\ \bar{\boldsymbol{r}}_2 = -Q_2^\top \boldsymbol{b}, \end{cases} \tag{3.12}$$

and (3.11) is equivalent to

$$\left\|-Q_2^\top \boldsymbol{b}\right\|_2^2 + \min_{\boldsymbol{x}\in\mathbb{R}^m} \left\{ \|\bar{\boldsymbol{r}}_1\|_2^2 \;\middle|\; \bar{\boldsymbol{r}}_1 = R_1 \boldsymbol{x} - Q_1^\top \boldsymbol{b} \right\}. \tag{3.13}$$

Note that $Q_1^\top \boldsymbol{b} = R_1 \boldsymbol{x}$ consist of m equations with m unknowns and can be solved by backward substitution since R_1 is an upper triangular matrix. A different way to solve these equation is by multiplying both side by R_1^{-1}: $\boldsymbol{x} = R_1^{-1} Q_1^\top \boldsymbol{b}$. Hence, the second part of (3.13) is equal to zero and therefore the sum of squared residuals is equal to

$$\|\bar{\boldsymbol{r}}\|_2^2 = \left\|Q_2^\top \boldsymbol{b}\right\|_2^2 . \tag{3.14}$$

This implies that if $\left\|Q_2^\top \boldsymbol{b}\right\|_2^2 = 0$ all equations $A\boldsymbol{x} = \boldsymbol{b}$ are satisfied. Moreover the minimizer of (3.11) is equal to

$$\boldsymbol{x} = R_1^{-1} Q_1^\top \boldsymbol{b}. \tag{3.15}$$

3.5. Iterative procedure

In Section 3.4, we included $\boldsymbol{x}^{(n-1)}$ in A. This suggests the following iterative procedure: choose an initial solution $\boldsymbol{x}^{(0)}$ and find a new solution $\boldsymbol{x}^{(1)}$ by solving the least square

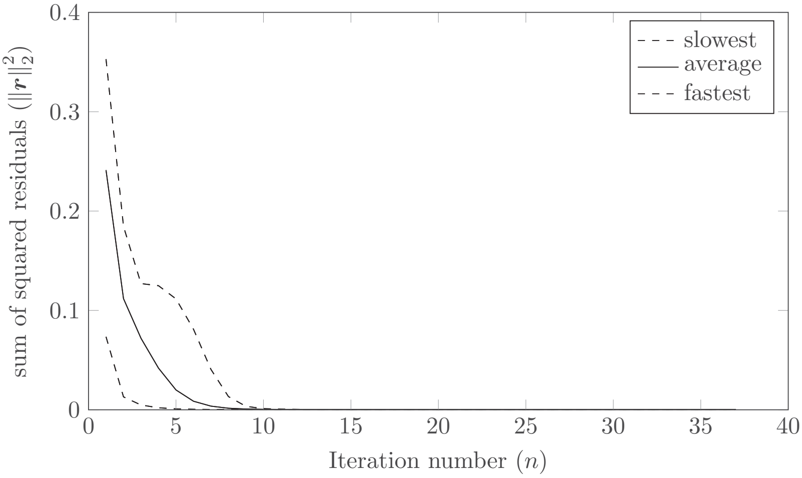


FIGURE 6. Convergence speed of the iterative procedure for the Rh blood group system.

problem described in the previous section. Hence, a solution to Eq. (3.9) can be found by iteratively solving

$$\boldsymbol{x}^{(n)} = R_1\left(\boldsymbol{x}^{(n-1)}\right)^{-1} Q_1\left(\boldsymbol{x}^{(n-1)}\right)^{\top}\boldsymbol{b}, \quad n = 1, 2, \ldots, \tag{3.16}$$

where the sum of squared residuals equals

$$\|\boldsymbol{r}\|_2^2 = \left\|Q_2\left(\boldsymbol{x}^{(n-1)}\right)^{\top}\boldsymbol{b}\right\|_2^2, \quad n = 1, 2, \ldots. \tag{3.17}$$

A solution $\boldsymbol{x}^{(n)}$ is defined satisfactory when $\|\boldsymbol{x}^{(n)} - \boldsymbol{x}^{(n-1)}\|_1 < 10^{-6}$. Hence, we consider (3.8) to be solved numerically.

3.6. Convergence of the algorithm

To support the iterative procedure from Section 3.5, we have performed numerical experiments for different blood group systems (i.e. ABO, Rhesus, Kell, Duffy, Kidd), phenotype distributions, and populations (e.g. European, American, African). We took $\boldsymbol{x}^{(0)} \in X_0$, where X_0 is the set of all identity vectors of size m. This means that every starting position $\boldsymbol{x}^{(0)}$ represents an initial population with only one genotype. Despite these extreme starting points all experiments converged to the same stationary solution $\boldsymbol{x}$ within 37 iterations. For the Rh system, which will be presented in Section 5.1 there are 36 starting points. In Figure 6, the fasted, slowest, and average convergence rates from these experiments for the Rh system are shown.

4. EFFECTIVENESS OF RECRUITING NEXT OF KIN FOR BLOOD DONORSHIP

The effectiveness of recruiting a next of kin for donorship for phenotype $\varphi \in \mathcal{F}$ is defined as the difference between the conditional probability that a relative of a donor has the same phenotype φ as the donor and the probability that an arbitrary individual in the population

Table 2. Probability that a relative of a donor has genotype γ_j given that the donor has genotype γ_i.

Relative	$\mathbb{P}[\text{relative } \gamma_j \mid \text{donor } \gamma_i]$
Sibling	$\displaystyle\sum_{\gamma_v \in \mathcal{G}} \sum_{\gamma_w \in \mathcal{G}} P(\gamma_j \mid \gamma_v, \gamma_w) \cdot \frac{P(\gamma_i \mid \gamma_v, \gamma_w) x_{\gamma_v} x_{\gamma_w}}{x_{\gamma_i}}$
Parent	$\displaystyle\sum_{\gamma_v \in \mathcal{G}} \frac{P(\gamma_i \mid \gamma_j, \gamma_v) x_{\gamma_j} x_{\gamma_v}}{x_{\gamma_i}}$
Child	$\displaystyle\sum_{\gamma_v \in \mathcal{G}} P(\gamma_j \mid \gamma_i, \gamma_v) x_{\gamma_v}$
Uncle / Aunt	$\displaystyle\sum_{\gamma_v \in \mathcal{G}} (\mathbb{P}[\text{sibling } \gamma_j \mid \text{parent } \gamma_v] \cdot \mathbb{P}[\text{parent } \gamma_v \mid \text{donor } \gamma_i])$
Nephew / Niece	$\displaystyle\sum_{\gamma_v \in \mathcal{G}} (\mathbb{P}[\text{child } \gamma_j \mid \text{sibling } \gamma_v] \cdot \mathbb{P}[\text{sibling } \gamma_v \mid \text{donor } \gamma_i])$
Grandparent	$\displaystyle\sum_{\gamma_v \in \mathcal{G}} (\mathbb{P}[\text{parent } \gamma_j \mid \text{parent } \gamma_v] \cdot \mathbb{P}[\text{parent } \gamma_v \mid \text{donor } \gamma_i])$

has this phenotype. That is

$$E_\varphi = \mathbb{P}\left[\text{relative } \varphi \mid \text{donor } \varphi\right] - \mathbb{P}\left[\text{individual } \varphi\right], \quad \forall \varphi \in \mathcal{F}. \tag{4.1}$$

We would like to rewrite this equation into the known phenotype distribution ($\boldsymbol{f}$), the heredity matrix (P), and the stationary genotype distribution ($\boldsymbol{x}$).

The conditional probability in Eq. (4.1) is, according to Bayes' rule, equal to

$$\begin{aligned}\mathbb{P}\left[\text{relative } \varphi \mid \text{donor } \varphi\right] &= \frac{\mathbb{P}\left[\text{donor } \varphi \ \cap \ \text{relative } \varphi\right]}{\mathbb{P}\left[\text{donor } \varphi\right]}, \\ &= \frac{1}{f_\varphi} \mathbb{P}\left[\text{donor } \varphi \ \cap \ \text{relative } \varphi\right].\end{aligned}$$

Since every genotype is related to a single phenotype the probability that the donor and its relative have the same phenotype can be computed by summing over all combinations of genotypes that they might have. Hence, if we denote the genotype of the donor by $\gamma_i \in \mathcal{G}$ and the genotype of its relative by $\gamma_j \in \mathcal{G}$, then we should sum over those combinations of genotypes for which $S_{i\varphi}$ and $S_{j\varphi}$ are both equal to one:

$$\mathbb{P}\left[\text{relative } \varphi \ \cap \ \text{donor } \varphi\right] = \sum_{\substack{\gamma_i \in \mathcal{G} \\ S_{i\varphi}=1}} \sum_{\substack{\gamma_j \in \mathcal{G} \\ S_{j\varphi}=1}} \mathbb{P}\left[\text{relative } \gamma_j \ \cap \ \text{donor } \gamma_i\right].$$

Moreover, applying Bayes' rule for the second time gives

$$\begin{aligned}\mathbb{P}\left[\text{relative } \varphi \mid \text{donor } \varphi\right] &= \frac{1}{f_\varphi} \sum_{\substack{\gamma_i \in \mathcal{G} \\ S_{i\varphi}=1}} \sum_{\substack{\gamma_j \in \mathcal{G} \\ S_{j\varphi}=1}} \left(\mathbb{P}\left[\text{relative } \gamma_j \mid \text{donor } \gamma_i\right] \cdot \mathbb{P}\left[\text{donor } \gamma_i\right]\right), \\ &= \frac{1}{f_\varphi} \sum_{\substack{\gamma_i \in \mathcal{G} \\ S_{i\varphi}=1}} \sum_{\substack{\gamma_j \in \mathcal{G} \\ S_{j\varphi}=1}} \left(\mathbb{P}\left[\text{relative } \gamma_j \mid \text{donor } \gamma_i\right] \cdot x_{\gamma_i}\right),\end{aligned}$$

where $\mathbb{P}[\text{relative } \gamma_j \mid \text{donor } \gamma_i]$ can be expressed in terms of P and $\boldsymbol{x}$ according to the relation between the donor and its relative, as is indicated in Table 2. The second probability in

Eq. (4.1) equals the frequency of φ in the general population. Hence, the effectiveness of recruiting a next of kin for donorship for phenotype $\varphi \in \mathcal{F}$ is equal to

$$E_{\varphi} = \frac{1}{f_{\varphi}} \sum_{\substack{\gamma_i \in \mathcal{G} \\ S_{i\varphi}=1}} \sum_{\substack{\gamma_j \in \mathcal{G} \\ S_{j\varphi}=1}} \left(\mathbb{P}\left[\text{relative } \gamma_j \mid \text{donor } \gamma_i\right] \cdot x_{\gamma_i}\right) - f_{\varphi}. \qquad \textbf{(4.2)}$$

5. APPLICATION TO MULTIPLE BLOOD GROUPS

In Section 2, we illustrated the effectiveness of a targeted donor recruitment strategy for siblings and parents of donors with particular ABO–RhD blood groups. To demonstrate the generic feature of our mathematical approach, we are going to analyze the more complicated Rh blood group system. Patients with sickle cell disease or thalassemia require regular (lifelong) blood transfusions. To prevent these recipients from forming antibodies against foreign RBC antigens, they are matched for a relatively large number of antigens. However, it is not easy to ensure that there is a sufficient number of required blood units available. We show how our generic model can be used to find more donors with the desired blood groups combinations.

5.1. Rh blood group system

In contrast to the well-known RhD blood group consisting of just three genotypes and two phenotypes, the full Rh system consists of 36 different genotypes and 18 different phenotypes. To compute the effectiveness of recruiting relatives of donors with desired Rh phenotypes, we first have to compute the stationary distribution based on Rh phenotype probabilities.

We computed these stationary probabilities by solving Eqs. (3.7) via an iterative procedure (3.16) (see Table 3). Next, we can apply use Bayes rule to compute the effectiveness of recruiting relatives as compared with random individuals from the general population. In Table 4 the calculated effectiveness of recruiting new donors among relatives is presented. Dependent of the family relationship effectiveness changes. Note that the effectiveness is variable, as it is dependent on the phenotype considered.

5.2. Recruitment of special blood groups

In the Netherlands, for some patient groups (e.g. women of reproductive age, patient with hemoglobinopathies) blood for transfusion is matched for up 13 antigens. With current recruitment strategies, it can be difficult to find enough donors with particular blood groups combinations. Moreover, due to the fact that the donor population in the Netherlands is mainly Caucasian, the patient population increasingly diversifying, and blood group frequencies differ between ethnic populations, the likelihood of finding suitable blood units for non-Caucasian individuals decreases. One of the main differences between phenotype frequencies of the Caucasian and African population is located in the so-called Duffy blood group system. This blood group system is similar to the ABO blood group system as it consists of six genotypes, four phenotypes $\mathcal{F} = (\text{Fy(a-b-)}, \text{Fy(a+b-)}, \text{Fy(a-b+)}, \text{Fy(a+b+)})$, and two antigens (Fy$^{\text{a}}$, and Fy$^{\text{b}}$). The phenotype frequencies for the Caucasian and African population are $\boldsymbol{f} = (0, 0.18, 0.33, 0.49)$ and $\boldsymbol{f} = (0.68, 0.06, 0.25, 0.01)$, respectively [13]. The probability that an African individual is Fy(a-b-) is 0.68 whereas the probability that a Caucasian individual has this combination is rare (< 0.0001). Hence, recruiting donors for

Table 3. Phenotype distribution of the Rh blood group system and the corresponding genotype distribution computed by the generic mathematical approach.

Phenotype	f	Genotype	x
CcDe	0.349	CDe/cde	0.326
		CDe/cDe	0.022
		cDe/Cde	0.001
CDe	0.185	CDe/CDe	0.176
		CDe/Cde	0.009
ce	0.151	CDE/cDE	0.151
CcDEe	0.133	CDe/cDE	0.119
		CDe/cdE	0.010
		cDE/Cde	0.003
		CDE/cde	0.002
		cDe/CdE	0.000
		CDE/cDe	0.000
cDEe	0.118	cDE/cde	0.110
		cDE/cDe	0.007
		cDe/cdE	0.001
cDE	0.023	cDE/cDE	0.020
		cDE/cdE	0.003
cDe	0.021	cDe/cde	0.020
		cDe/cDe	0.001
cEe	0.009	cdE/cde	0.009
Cce	0.008	Cde/cde	0.008
CDEe	0.002	CDE/CDe	0.002
		CDE/Cde	0.000
CcDE	0.001	CDE/cDE	0.001
		CDE/cdE	0.000
Other	0.000		

Table 4. Effectiveness of recruiting new donors among relatives of donors with a known Rh phenotype. Although this effectiveness depends on the specific phenotype sought for, we reported a minimum, average, and maximum effectiveness measure.

	Effectiveness		
Relative	Minimum	Average	Maximum
Sibling	0.197	0.275	0.361
Parent	0.068	0.156	0.245
Child	0.068	0.156	0.245
Uncle / Aunt	0.028	0.078	0.122
Nephew / Niece	0.028	0.078	0.122
Grandparent	0.034	0.078	0.122

this specific blood group randomly within the Caucasian population is virtually impossible. However, for a donor with this blood group combination, the probability that a sibling has the same combination is 25% if the donor is Caucasian, and 83% if the donor is African.

6. CONCLUSIONS

The generic mathematical approach described in this paper allows computing a stationary genotype distribution for a given set of blood groups, which may even belong to multiple blood group systems. The input for the model consists of the phenotype distributions in a population only. This stationary genotype distribution allows answering a number of interesting questions using elementary statistics.

This paper was tailored to quantify the effectiveness of targeted recruitment strategies aiming for relatives of donors with specific blood groups. It shows that the impact, in terms of the efficiency of targeting the next of kin of donors with known blood groups as potential new donors, can be substantial.

Recently, another application was found in computing the probability of a blood group mismatch between mother and fetus during pregnancy. This analysis also required an estimate of the stationary distribution of genotypes as a basis for further calculations. The approach outlined in this paper, therefore, seems promising for answering various question related to genetic counseling.

Acknowledgements

This work was performed as part of the BloodMatch project, which is a research project supported by Sanquin Blood Supply Foundation (PPOC-14-25). The authors are most grateful to both reviewers and the associate editor from which this paper has well benefited.

References

1. Bernstein, S. (1924). Solution of a mathematical problem related to the theory of inheritance. *Uch. Zap. n.-i. kaf. Ukrainy* 1: 83–115.
2. Brousseau, D.C., Panepinto, A.J., Nimmer, M., & Hoffmann, R.G. (2010). The number of people with sickle-cell disease in the united states: national and state estimates. *American Journal of Hematology* 85(1): 77–78.
3. Carter, M.C., Wilson, J., Redpath, G.S., Hayes, P., & Mitchell, C. (2011). Donor recruitment in the 21st century: challenges and lessons learned in the first decade. *Transfusion and Apheresis Science* 45(1): 31–43.
4. Casas, J., Ladra, M., Omirov, B., & Turdibaev, R. (2016). On the algebraic properties of the human abo-blood group inheritance pattern. *The ANZIAM Journal* 58(1): 78–95.
5. Davey, R.J. (2004). Recruiting blood donors: challenges and opportunities. *Transfusion* 44(4): 597–600.
6. Elston, R.C., Satagopan, J.M., & Sun, S. (2012). Genetic terminology. In Statistical Human Genetics. New York, NY, USA: Humana Press, pp. 1–9.
7. Ganikhodjaev, N., Daoud, J., & Usmanova, M. (2010a). Linear and nonlinear models of heredity for blood groups and rhesus factor. *Journal of Applied Sciences* 10(16): 1748–1754.
8. Ganikhodjaev, N., Jamal Ibrahim, D., & Usmanova, M. (2010b). Stochastic models of heredity rhesus factor. *Australian Journal of Basic and Applied Sciences* 4(8): 3306–3310.
9. Ganikhodjaev, N., Saburov, M., & Jamilov, U. (2013). Mendelian and non-mendelian quadratic operators. *Applied Mathematics & Information Sciences* 7(5): 1721–1729.
10. Ganikhodzhaev, R., Mukhamedov, F., & Rozikov, U. (2011). Quadratic stochastic operators and processes: results and open problems. *Infinite Dimensional Analysis, Quantum Probability and Related Topics* 14(2): 279–335.
11. Guo, S.W. & Thompson, E.A. (1992). Performing the exact test of hardy-weinberg proportion for multiple alleles. *Biometrics* 48(2): 361–372.
12. International Society of Blood Transfusion (2014). Table of blood group systems v4.0. http://www.isbtweb.org/working-parties/red-cell-immunogenetics-and-blood-group-terminology/ [Accessed: January 2017].
13. Reid, M.E., Lomas-Francis, C., & Olsson, M.L. (2012). The blood group antigen factsbook. Waltham, Massachusetts, USA: Academic Press.
14. Sadykov, T. (2017). Polynomial dynamics of human blood genotypes frequencies. *Journal of Symbolic Computation* 79: 342–355.

15. Trefethen, L.N. & Bau III, D. (1997). Numerical linear algebra. Philadelphia, Pennsylvania, USA: Siam.
16. Williamson, L.M. & Devine, D.V. (2013). Challenges in the management of the blood supply. *The Lancet* 381(9880): 1866–1875.

APPENDIX A

Sets

$\mathcal{A}$	Set of antigens,
$\mathcal{L}$	Set of alleles,
$\mathcal{H}$	Set of haplotypes (index h),
$\mathcal{G}$	Set of genotypes, with cardinality $\lvert\mathcal{G}\rvert = m$ (index γ),
$\mathcal{F}$	Set of phenotype, with cardinality $\lvert\mathcal{F}\rvert = n$ (index φ).

Matrices

$S \in \{0,1\}^{m\times n}$	Matrix mapping blood group genotypes to blood group phenotypes,
$P \in \mathbb{R}^{m\times m\times m}$	Heredity matrix, with $P(\gamma_k \mid \gamma_i, \gamma_j)$ the probability that two parents with genotypes γ_i and γ_j conceive a child with genotype γ_k,
$A \in \mathbb{R}^{(m+n)\times m}$	Matrix used for the linearization of a system of quadratic equations, $(A^\top = [\boldsymbol{x}^\top P - I \qquad S^\top])$,
$I \in \mathbb{R}^{m\times m}$	Identity matrix,
$Q \in \mathbb{R}^{(m+n)\times(m+n)}$	Orthogonal matrix used for QR factorization, $Q = [Q_1 \quad Q_2]$, $Q_1 \in \mathbb{R}^{(m+n)\times m}$, $Q_2 \in \mathbb{R}^{(m+n)\times n}$,
$R \in \mathbb{R}^{(m+n)\times m}$	Upper triangular matrix used for the QR decomposition, $R^\top = [R_1 \quad 0]$, $R_1 \in \mathbb{R}^{m\times m}$, $0 \in \mathbb{R}^{n\times m}$.

Vectors

$\boldsymbol{x} \in \mathbb{R}^{m\times 1}$	Column vector representing a stationary genotype distribution,
$\boldsymbol{f} \in \mathbb{R}^{n\times 1}$	Column vector representing the distribution of phenotypes in the general population,
$\boldsymbol{v}_h \in \mathbb{R}^{m\times 1}$	Column vector containing the probabilities that a parent transmits haplotype $h \in \mathcal{H}$ to the child,
$\boldsymbol{b} \in \mathbb{R}^{(m+n)\times 1}$	Column vector used for the linearizion of a system of quadratic equations $(\boldsymbol{b}^\top = [\boldsymbol{0} \quad \boldsymbol{f}])$,
$\boldsymbol{r} \in \mathbb{R}^{(m+n)\times 1}$	Column vector for the residuals $\boldsymbol{r} = A\boldsymbol{x} - \boldsymbol{b}$.

The symbols $\mathcal{H}'$, $\mathcal{G}'$, B, $\boldsymbol{e}_m$, $\boldsymbol{\pi}$, $\boldsymbol{y}$ and a are omitted here, as they will be defined and only used in Appendix C.

APPENDIX B

This appendix provides a calculation for the conditional probability that a sibling of a donor has phenotype D, given that it is known that this donor has phenotype D, or mathematically stated:

$$\mathbb{P}\left[\text{sibling D} \mid \text{donor D}\right].$$

First Bayes' rules is applied giving

$$= \frac{\mathbb{P}\left[\text{sibling D} \cap \text{donor D}\right]}{\mathbb{P}\left[\text{donor D}\right]}.$$

Then we replace the numerator by enumerating over all possible combinations of genotypes that the donor and its sibling might have if the both have phenotype D:

$$\begin{aligned} &= \frac{1}{\mathbb{P}\left[\text{donor D}\right]} \cdot \big(\mathbb{P}\left[\text{sibling } DD \cap \text{donor } DD\right] + \mathbb{P}\left[\text{sibling } DD \cap \text{donor } Dd\right] \\ &\quad + \mathbb{P}\left[\text{sibling } Dd \cap \text{donor } DD\right] + \mathbb{P}\left[\text{sibling } Dd \cap \text{donor } Dd\right]\big). \end{aligned}$$

Subsequently, enumerate over all possible combinations of genotypes that the parents might have:

$$\begin{aligned} &= \frac{1}{\mathbb{P}\left[\text{donor D}\right]} \cdot \sum_{\gamma_v \in \mathcal{G}} \sum_{\gamma_w \in \mathcal{G}} \Big(\big(\mathbb{P}\left[\text{sibling } DD \cap \text{donor } DD \mid \text{parents } (\gamma_v, \gamma_w)\right] \\ &\quad + \mathbb{P}\left[\text{sibling } DD \cap \text{donor } Dd \mid \text{parents } (\gamma_v, \gamma_w)\right] \\ &\quad + \mathbb{P}\left[\text{sibling } Dd \cap \text{donor } DD \mid \text{parents } (\gamma_v, \gamma_w)\right] \\ &\quad + \mathbb{P}\left[\text{sibling } Dd \cap \text{donor } Dd \mid \text{parents } (\gamma_v, \gamma_w)\right]\big) \cdot \mathbb{P}\left[\text{parents } (\gamma_v, \gamma_w)\right]\Big). \end{aligned}$$

Rewriting the above formula gives:

$$\begin{aligned} &= \frac{1}{f_D} \cdot \sum_{\gamma_v \in \mathcal{G}} \sum_{\gamma_w \in \mathcal{G}} \Big(\big(P\left(DD \mid \gamma_v, \gamma_w\right) \cdot P\left(DD \mid \gamma_v, \gamma_w\right) + P\left(DD \mid \gamma_v, \gamma_w\right) \cdot P\left(Dd \mid \gamma_v, \gamma_w\right) \\ &\quad + P\left(DD \mid \gamma_v, \gamma_w\right) \cdot P\left(Dd \mid \gamma_v, \gamma_w\right) + P\left(Dd \mid \gamma_v, \gamma_w\right) \cdot P\left(Dd \mid \gamma_v, \gamma_w\right)\big) \cdot x_{\gamma_v} x_{\gamma_w}\Big) \\ &= \frac{1}{f_D} \cdot \sum_{\gamma_v \in \mathcal{G}} \sum_{\gamma_w \in \mathcal{G}} \Big(\left(P\left(DD \mid \gamma_v, \gamma_w\right) + P\left(Dd \mid \gamma_v, \gamma_w\right)\right)^2 \cdot x_{\gamma_v} x_{\gamma_w}\Big). \end{aligned}$$

Filling in the numbers presented in the table below gives us the following solution for the probability that the sibling of the donor and the donor have both a RhD-pos phenotype:

$$\begin{aligned} &= \frac{1}{0.850} \cdot (0.141 + 0.178 + 0.056 + 0.178 + 0.127 + 0.018 + 0.056 + 0.018 + 0.000) \\ &= 0.908. \end{aligned}$$

v	w	$(P(DD \mid \gamma_v, \gamma_w) + P(Dd \mid \gamma_v, \gamma_w))^2$	x_{γ_v}	x_{γ_w}
DD	DD	$(1+0)^2$	0.375	0.375
DD	Dd	$(\frac{1}{2}+\frac{1}{2})^2$	0.375	0.475
DD	dd	$(0+1)^2$	0.375	0.150
Dd	DD	$(\frac{1}{2}+\frac{1}{2})^2$	0.475	0.375
Dd	Dd	$(\frac{1}{4}+\frac{1}{2})^2$	0.475	0.475
Dd	dd	$(0+\frac{1}{2})^2$	0.475	0.150
dd	DD	$(0+1)^2$	0.150	0.375
dd	Dd	$(0+\frac{1}{2})^2$	0.150	0.475
dd	dd	$(0+0)^2$	0.150	0.150

APPENDIX C

In this entire appendix, we fix ourselves to one and the same male genotype distribution $\boldsymbol{x}$ (as to be conceived for an iteration step as $\boldsymbol{x} = \boldsymbol{x}^{(n-1)}$). We only include $\boldsymbol{x}$ in the representation of $\mathcal{H}'(\boldsymbol{x})$ and $\mathcal{G}'(\boldsymbol{x})$ as these can be explicitly different for different $\boldsymbol{x}$.

Now, for given $\boldsymbol{x}$, let $A \in \mathbb{R}^{(m+n)\times m}$ be defined as follows:

$$A = \begin{bmatrix} \boldsymbol{x}^\top P - I \\ S^\top \end{bmatrix},$$

where $\boldsymbol{x} \in \mathbb{R}^{m\times 1}$ is the genotype distribution, $P \in \mathbb{R}^{m\times m\times m}$ is the heredity matrix, $I \in \mathbb{R}^{m\times m}$ is the identity matrix, and $S \in \{0,1\}^{m\times n}$ is the matrix mapping blood group genotypes to blood group phenotypes. It need to be shown that A has full column rank. Let $B \in \mathbb{R}^{m\times m}$ be a matrix ($B^\top = \boldsymbol{x}^\top P$), with

$$B_{jk} = \sum_{\gamma_i \in \mathcal{G}} P\left(\gamma_k \mid \gamma_i, \gamma_j\right) x_{\gamma_i},$$

where B_{jk} can be interpreted as the probability that a mother with genotype γ_j conceives a child with genotype γ_k, given that the genotype distribution of the father equals $\boldsymbol{x}$. Clearly, B is a stochastic matrix.

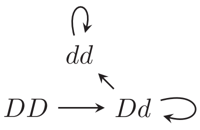
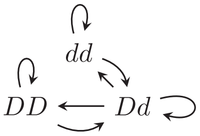
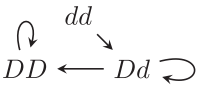
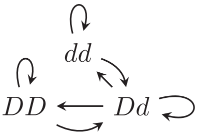
We are now going to use B as a transition matrix of a Markov chain with state space $\mathcal{G}$. Hence, purely abstractly, the outcomes of the Markov chain can be seen as keeping track of the genotype distribution of mothers, which are based on the heredity matrix P and a *fixed* genotype distribution $\boldsymbol{x}$. (For the simple RhD example, Table 5 shows Markov chains for different $\boldsymbol{x}$.)

Let $\mathcal{H}'(\boldsymbol{x}) \subseteq \mathcal{H}$ be defined as the set of haplotypes that a father may transmit to his child, or differently stated: for every haplotype $h \in \mathcal{H}'(\boldsymbol{x})$ there exists a genotype $\gamma_i \in \mathcal{G}$, $\gamma_i = \{h, \cdot\}$ and/or $\gamma_i = \{\cdot, h\}$, such that $x_{\gamma_i} > 0$. This implies that the following two conditions should be satisfied:

$$\begin{cases} \displaystyle\sum_{\gamma_i \in \mathcal{G}: h \in \gamma_i} x_{\gamma_i} > 0 & \text{if } h \in \mathcal{H}'(\boldsymbol{x}), \\ \displaystyle\sum_{\gamma_i \in \mathcal{G}: h \in \gamma_i} x_{\gamma_i} = 0 & \text{if } h \notin \mathcal{H}'(\boldsymbol{x}). \end{cases}$$

As a consequence, roughly speaking, the haplotypes that are not present in the male population will eventually disappear. More precisely, the Markov chain, depending on the given genotype

TABLE 5. Markov chains for the RhD example.

$\boldsymbol{x}$	$\begin{array}{ccc} dd & Dd & DD \\ [\,1 & 0 & 0\,] \end{array}$	$\begin{array}{ccc} dd & Dd & DD \\ [\,0 & 1 & 0\,] \end{array}$	$\begin{array}{ccc} dd & Dd & DD \\ [\,0 & 0 & 1\,] \end{array}$	$\begin{array}{ccc} dd & Dd & DD \\ [\,\frac{1}{2} & 0 & \frac{1}{2}\,] \end{array}$
B	$\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$
Markov chain				
$\mathcal{H}'(\boldsymbol{x})$	{d}	$\{d, D\}$	{D}	$\{d, D\}$
$\mathcal{G}'(\boldsymbol{x})$	$\{dd\}$	$\{dd, Dd, DD\}$	$\{DD\}$	$\{dd, Dd, DD\}$

distribution $\boldsymbol{x}$, will always have a single closed class given by

$$\mathcal{G}'(\boldsymbol{x}) = \left\{\gamma_k \in \mathcal{G} \mid \gamma_k = \{h_{k_1}, h_{k_2}\}, \quad h_{k_1}, h_{k_2} \in \mathcal{H}'(\boldsymbol{x})\right\}.$$

Note that for $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, two *fixed* genotype distributions, and $a \in (0,1)$ the following relation holds $\mathcal{H}'(a\boldsymbol{x}_1 + (1-a)\boldsymbol{x}_2) = \mathcal{H}'(\boldsymbol{x}_1) \cup \mathcal{H}'(\boldsymbol{x}_2)$, but $\mathcal{G}'(a\boldsymbol{x}_1 + (1-a)\boldsymbol{x}_2) = \mathcal{G}'(\boldsymbol{x}_1) \cup \mathcal{G}'(\boldsymbol{x}_2)$ is generally not true (see also Table 5).

Since the Markov chain has a single closed class, there exists a unique stationary distribution π satisfying

$$\begin{cases} B^\top \boldsymbol{\pi} = \boldsymbol{\pi} \\ \boldsymbol{e}_m^\top \boldsymbol{\pi} = 1 \end{cases} \qquad \Rightarrow \quad \begin{bmatrix} B^\top - I \\ \boldsymbol{e}_m^\top \end{bmatrix} \boldsymbol{\pi} = \begin{bmatrix} \boldsymbol{0} \\ 1 \end{bmatrix},$$

where $\boldsymbol{e}_m$ is the all ones vector of length m. What remains, is to use this statement to proof the full column rank property of A.

The only candidate vector $\boldsymbol{y}$ (up to a constant), such that $A\boldsymbol{y} = \boldsymbol{0}$, is $\boldsymbol{y} = \boldsymbol{\pi}$, the stationary distribution of the matrix B. Now note that column sum of each column of S is at least one (i.e., $S^\top \boldsymbol{e}_n \geq \boldsymbol{e}_m$) and therefore $S^\top \boldsymbol{y} = \boldsymbol{0}$ if $\boldsymbol{y} = \boldsymbol{0}$. Hence, the only vector satisfying $A\boldsymbol{y} = \boldsymbol{0}$ is $\boldsymbol{y} = \boldsymbol{0}$, which implies that the matrix A has full column rank.