Journal of Emerging Computer Technologies
Research Article
Received: *2023-10-17* | ***Reviewing:*** *2023-10-19* & *2024-01-07*| ***Accepted:*** *2024-01-08* | ***Online:*** *2024-01-08* | ***Issue Date:*** *2023-12-31*
Doi: 10.57020/ject.1375605

A Dynamic Method and Program for Disease-Based Genetic Classification of Individuals

Onur Çakırgöz
Department of Computer Engineering
Bartın University
Bartın, Turkey
onurcakirgoz@bartin.edu.tr
0000-0002-9347-1105

Süleyman Sevinç
Labenko Bilişim A.Ş.
İzmir, Turkey
suleysevinc@gmail.com
0000-0001-9052-5836

***Abstract*—Personalized medicine is gaining increasing importance. However, genetic-based diseases have different underlying genetic factors, requiring separate relative risk models for each disease. In addition to these difficulties, comparing individuals according to their genetic characteristics and determining a personalized treatment method based on this, is a separate problem which is very difficult to do manually. In this study, a dynamic classification method and program is proposed for disease-based classification of individuals according to their genetic characteristics. To the best of our knowledge, this is the first generic method which performs disease-based classification of individuals. In the developed program, relative risk models containing only genetic factors are an input of the program and a common format has been created for this purpose. Our generic classification method classifies people by using information from any relative risk model rearranged according to the common format. Thanks to this program, relative risk models can be managed from a single point, many people can be classified based on their genetic characteristics, and individuals, who are genetically most similar to any person, can be determined by experts using the outputs (relevant tables) of the program.**

Keywords—personalized medicine, computational medicine, genetic classification, relative risk model, genetic similarity, genetic variation

I. INTRODUCTION

Many common human diseases and traits are affected by several genetic and environmental factors [1, 2]. To investigate the genetic variants contributing to these human diseases, researchers do candidate gene studies and Genome-wide association studies (GWAS) [3-5]. Until today, GWAS and gene studies have determined many variants associated with diseases and have provided so many relative risk models [3, 6]. As known, there are so many genetic diseases, and these diseases can be classified as single gene mendelian diseases and complex diseases [7, 8]. In single gene mendelian diseases, as is evident from its name, the variant or variants associated with the disease present at one gene. Additionally, the number of variants associated with the disease is relatively too small [7]. Clinicians can easily diagnose the single gene mendelian diseases most of the time, but unfortunately the things they can do medically are limited [9]. On the other hand, in complex diseases, there might be tens of disease-associated variants [10-12]. For instance, the number of genetic variants contributing to the Crohn's disease is 32 [12]. On the other hand, the number of variants associated with the diseases such as age-related macular degeneration [13], type 2 diabetes [14, 15], early onset myocardial infarction [16] are 5, 18 and 9, respectively. Contrary to the single gene mendelian diseases, clinicians often have more opportunities in complex diseases [4, 17].

There are many genetic-based complex diseases and the numbers of variations associated with these diseases vary considerably. Beyond the number of variations, the properties of the variations vary as well [3, 4, 8]. Most of the variations associated with the diseases increase the risk of developing the disease, whereas others decrease the risk (protective). Sometimes reference allele has high risk; and sometimes alternate allele has high risk. On the other hand, the genetic characteristics of the diseases or traits might differ from population to population or from region to region [18]. A remarkable information is that the vast majority of GWAS and other genetic studies have been limited to European ancestry populations [3,7, 19, 20]. Fewer studies have been carried out in non-European countries (especially populations of under-developed or developing countries) and these studies have determined intriguing new variants. Due to the economic reasons, however, the physicians serving in under-developed or developing countries might have to use the relative risk model generated for a different population (at least until a genome-wide association study is made in its own population for that disease).

With the understanding of the effect of genetic factors, personalized medicine concept has entered our lives [21-24]. Unlike traditional medicine, personalized medicine has adopted the person-specific treatment approach [21, 22]. At this point, the most important mainstay of personalized medicine is genetic characteristics and genetic variations [24, 25]. Since a substantial portion of individual differences in the predisposition to complex disease is due to genetic variants, identifying these variants provides better prevention, diagnosis, and treatment of disease [26]. Within the scope of personalized medicine approach, physicians can inform individuals as to how they can behave to prevent the disease [24]. Furthermore, the individuals who caught a disease and have undergone a successful treatment process, their personal genotypes and the applied treatment method can be stored in a database. When a new patient who has contracted the same disease is admitted to the clinic, clinicians can use the system to identify patients who are genetically most similar to that patient. They can then apply a previously successful treatment method. Unfortunately, to perform these medical applications, a dynamic program is needed that can automatically classify individuals based on their genetic characteristics (disease-

Cite (APA) : Çakırgöz, O., Sevinç, S. (2023). A Dynamic Method and Program for Disease-Based Genetic Classification of Individuals. *Journal of Emerging Computer Technologies*, 3(1), 12-20. Doi: 10.57020/ject.1375605

based) and support relative risk models. Based on this, in this study, a disease-based genetic classification method was designed. This method supports all types of genetic variations. Additionally, a dynamic desktop application was developed. This application utilizes the aforementioned classification method and also supports relative risk models.

In bioinformatics field, after the Human Genome Project [27, 28], the International HapMap Project [29] and 1000 Genome Project [30-33] are considered important turning points. These projects, which catalog the human genetic variations, are put in a separate place. The variation data of 2504 anonymous people were published by 1000 Genomes Project as VCF [34] and BCF [35] files. Within this study, the personal variation data of 2504 people published by 1000 Genomes Project were used.

The sections of the article are as follows: In the second section, firstly, the disease-based genetic similarity approach is mentioned, and the basic classification method is introduced, then, the common format created for the relative risk models is explained, and finally, the developed dynamic program and the components that make up this program are explained. In the third chapter, after loading the sample data to the program, the results produced by the program are shown and these results are discussed. Final remarks are given in the last section.

II. MATERIALS AND METHODS

A. Disease-Based Genetic Similarity and The Basic Classification Method

To be able to perform personalized medicine applications, finding the genetic similarities and dissimilarities of individuals between each other is sometimes required. Considering that, human DNA consists of 3.2 billion base-pairs, humans are diploid and millions of variations exist in the genome of each person, it is seen that the concept of genetic similarity is actually a very complex and broad subject; but what is meant here is disease-based genetic similarity (or classification). As is known, genome-wide association studies aim to identify the variations that influence the disease and to what extent they affect it, by evaluating a number of variations. Thereby, since one of the main objectives is to find genetically most similar individuals to an individual who caught a particular disease, numerous variations which are not associated with that disease should not be considered in genetic similarity between individuals. Besides, since the variations associated with each disease differ, genetic similarity between individuals must be found separately for each disease. As a result, relative risk models generated by genome-wide association studies should be used as the basis for disease-based classification. In parallel, the risk estimations and groups (genotype combinations) specified in the relative risk model are used in our classification method.

The genetic elements that reveal diseases are variations (single nucleotide polymorphisms, short insertions and deletions, and structural variants) occurring in specific locations. At the level of phenotype, another important factor which may determine the emergence of the disease is the presence or absence of genetic variations in both alleles of the person. On the other hand, as we know, the allele counts (reference allele + alternate alleles) of the variations vary. For example, while any variation may have only 2 alleles, another variation may have 3-4 alleles. Given that the human genome is diploid, the number of possible genotypes naturally varies. Depending on the number of alleles of the variation, the number of genotypes that can form is determined according to (1). Thus, when we want to classify individuals based on a specific variation that leads to disease, we can obtain the class count through (1).

$$G_{Nd} = A_{Nd}{}^{2} - \binom{A_{Nd}}{2} \quad (1)$$

The symbols used in this equation and in subsequent equations are shown in Table 1.

TABLE I. SYMBOL TABLE FOR THE DISEASE-BASED CLASSIFICATION APPROACH

Symbol	Explanation
i	Class number
K_d	In how many steps the genotypes of the d^{th} variation vary, that is, the step count of the d^{th} variation
G_{Nd}	Number of genotypes of d'th variation associated with the disease
A_{Nd}	Number of alleles of d'th variation associated with the disease
AV_N	Number of variations associated with the disease
G_{CCd}	The column corresponding to the genotype – Numerical equivalent of the genotype regarding the d^{th} variation
A_{LI}	The index of the lower index allele constituting the genotype
A_{HI}	The index of the higher index allele constituting the genotype
PC_N	Number of possible classes associated with the disease
N_{Nd}	Number of nodes at depth d of the classification tree
UN_{Nd}	Number of unnecessary nodes at depth d of the classification tree
UN_{TN}	Total number of unnecessary nodes on the classification tree

In fact, the number of genotypes that can occur is the square of the number of alleles. But, since we consider genotypes from a functional standpoint and (most of the time) half of the heterozygous genotypes are functionally equivalent to the other half, we subtract half of the heterozygous genotypes from the square of the number of alleles. (If heterozygous genotypes are not functionally equivalent, (1) should be updated accordingly.) For example, suppose that the number of alleles of a variation is two and these are 'A' (reference) and 'T' (alternate and haplotype increasing disease risk) bases. In this case, the genotypes that can be formed are "AA", "TT", "AT" and "TA". Since "AT" and "TA" are functionally equivalent, both are accepted as one. The resulting classes can be defined as: the variation does not exist in both alleles ("AA", class-1), exists only in one allele ("AT" or "TA", class-2), and exist in both alleles ("TT", class-3).

Genetic similarity relationship between classes is a process beyond that and calculating genetic similarities between classes is beyond the scope of this study. Describing and determining the genetic similarity relationship clearly between the resulting genotypes (classes) is not easy most of the time. Since the ultimate goal of doctors is to apply a previously successful personalized treatment to a similar patient, different parameters such as drug-protein interaction may also come into play in the genetic similarity relationship here. In this respect, determining the disease-based genetic similarity relationship is a process that should be done after classification, sometimes it depends on different parameters and therefore, the main decision maker is the doctors (experts). The generic classification method and program proposed in this study systematically presents the relevant classification results to the users. If it is assumed that there are no different parameters in the genetic similarity relationship between classes in the above example, the similarity relationship can be expressed as: The class which is genetically most similar to both classes 1 and 3, is class 2. To find the class, which is genetically most similar to class 2, it is required to look at the risk estimations specified in the relative risk model of the disease. For instance, assume that the classes' disease risks are specified as 1, 3, and 9 in the model, respectively; in that case, due to the lower risk difference, class 1 is the genetically most similar class to class 2. Therefore, sometimes the relative risk model can provide insight into the disease-based genetic similarity relationship.

Genome-wide association studies have shown that the number of susceptibility variants associated with any disease is more than one most of the time. Therefore, in these cases, classification depicted in the previous example is not sufficient. In the case that the number of susceptibility variants associated with the disease is more than one, the resulting number of classes can be obtained using (2). In our method, each different combination of genotypes corresponds to a different class.

$$\boldsymbol{PC_N = \prod_{d=0}^{AV_N-1} G_{Nd}} \tag{2}$$

When viewed from the perspective of computer science, the image that emerges from the classification approach of humans depending on the variations associated with the disease or treat is a tree. In other words, for the classification of individuals based on their genetic characteristics, tree is the first data structure which comes to mind. Of course, this tree data structure would be specific to the problem of classification of people according to their genetic characteristics. As a result, the nodes of this tree are variations, and the branches are genotypes. On the other hand, any path on the tree corresponds to a class (the genotype characteristics of the class), and the individuals involved in a class can be thought as a leaf of the tree. The tree method might seem like an efficient way at first glance, but all the nodes at the same depth have the same value (variation). This means that this tree uses redundant nodes. In addition, the number of unnecessary nodes in each level increases exponentially, depending on the depth of that level. Equation (3) is used to find the number of nodes at any depth of the disease-based classification tree.

$$\boldsymbol{N_{Nd} = \begin{cases} 1, & d == 0 \\ \prod_{k=0}^{d-1} G_{Nk}, & d > 0 \end{cases}} \tag{3}$$

As we have already stated, the values of all nodes at a particular depth of this tree are the same. Actually, only one of these nodes is enough. From this point of view, we can use (4) to calculate the number of unnecessary nodes at a particular depth.

$$\boldsymbol{UN_{Nd} = N_{Nd} - 1} \tag{4}$$

Equation (4) calculates the number of redundant nodes only at a certain depth of the tree. If we want to calculate the number of all unnecessary nodes in the tree, the solution is quite simple. For this, it is necessary to sum the number of unnecessary nodes in each level of the tree. Accordingly, Equation (5) was developed to acquire the total number of redundant nodes on the tree.

$$\boldsymbol{UN_{TN} = \sum_{d=0}^{AV_N-1} UN_{Nd}} \tag{5}$$

The situation that the disease-based classification tree uses redundant nodes and the high cost of the search operation required to find out which class an individual belongs to, have directed us to another data structure. At the beginning, we noticed that the nodes (variations) and the branches (genotypes) of the tree can be stored in a table. In this way, all the variations are stored only once in the table. For instance, when the number of susceptibility variations associated with the disease is 2, the resulting table is seen in Table 2. In this example, the number of alleles (haplotypes) of both variations is two and the respective alleles are indicated in the table. As seen from the table, the number of classes for this disease is found as 9 according to (2), and these classes are numbered from 0 to 8. On the two lines below the class numbers, the corresponding genotypes to that class appear.

TABLE II. CLASSES AND THEIR GENOTYPE PROPERTIES IN THE CASE THAT $AV_N = 2$

Variations	**Classes and Corresponding Genotypes**								
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
d=0 (Alleles=G,C)	G|G	G|G	G|G	G|C	G|C	G|C	C|C	C|C	C|C
d=1 (Alleles=T,A)	T|T	T|A	A|A	T|T	T|A	A|A	T|T	T|A	A|A

The data in Table 2 can be stored in a two-dimensional array. As a matter of fact, our first thought was in this direction as well. Afterwards, while doing operations on the table, we realized that the genotype combinations actually form a pattern, a connection can be established between the classes and the genotypes by using this pattern, and therefore, it is unnecessary to keep the data in Table 2 in any data structure. As can be seen from Table 2, classes are represented by numbers. On the other hand, genotypes are represented by nucleotide bases (letters). In addition, since the alleles of each variation vary, naturally, genotypes also vary. In fact, the important point here is the placement of the genotypes on the table. The genotypes are located in a certain order on the table. For instance, the genotypes of the first variation vary in every 3 steps. On the other hand, this number is 1 for the second variation, and when the genotypes of both variations are

considered together, the resulting genotype combinations form a pattern. Thanks to the pattern that the genotype combinations form, a link can be established between the classes and the genotypes, but the difficulty here is that the genotypes are represented by nucleotide bases. The formulas that will establish the link between genotypes and classes need numerical data. Therefore, primarily, we need to convert the genotypes into numbers. The numerical equivalents of the genotypes in Table 2 are shown in Table 3.

TABLE III. THE NUMERICAL EQUIVALENTS OF THE GENOTYPES IN TABLE 2

Variations	**Classes and Corresponding Genotypes**								
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
d=0 (Alleles=G,C)	0	0	0	1	1	1	2	2	2
d=1 (Alleles=T,A)	0	1	2	0	1	2	0	1	2

Regardless of the number of alleles of the variation, the genotypes that are formed need to be converted into numbers automatically. In other words, there is a need for an equation for the process of converting the genotype to number. For this purpose, (6) was developed. At this point, we must specify that the alleles constituting the genotype are located on the respective "Alleles" array and the A_{LI} and A_{HI} variables in (6) take the values of the indices of these alleles. "Alleles" arrays are the important attributes found in the Variant array, and the Variant array will be described in the section "The Dynamic Program and Its Components".

$$\boldsymbol{G_{CCd} = \left(\sum_{j=0}^{A_{LI}} (A_{Nd} - j)\right) - (A_{Nd} - A_{HI})} \quad (6)$$

The following "Convert_into_genotype(index)" function has been developed to transform the integer back to the genotypes. The "Convert_into_genotype(index)" function is the inverse of (6). Since it is not possible to express this transformation in the form of equation, only the pseudo-code of the function is given here. The variable named "A" in this function represents the "Alleles" attribute explained in the previous paragraph.

```
Convert_into_genotype(index)
 1 Let alleles be a string array of length 2
 2 Let temp and temp2 be integers
 3 temp = index;
 4 for(i = 0; i < A.Length; i++)
 5     temp2 = temp;
 6     temp -= (A.Length-i);
 7     if (temp >= 0)
 8         continue;
 9     else
10         alleles[0] = A[i];
11         alleles[1] = A[i + (temp2 % (A.Length - i))];
12         break;
13 return alleles;
```

In the above section, mainly the conversion of the genotype to the number and the conversion of the number to the genotype were described; namely, the processes and equations related to a single genotype were mentioned. From this point on, the formulas necessary for the relation between the genotypes (number equivalent) and the classes will be addressed. First of all, the relation between the classes and the genotypes is bidirectional. Namely, the genotypes of any class whose number is specified can be determined, or vice versa. Equation (7), which was developed for the first direction of the relation, is below. Thanks to this equation, the genotype (numerical equivalent) corresponding to the respective level (d^{th} variation) of any class, whose number ("i") is specified can be easily found. The "d" parameter of this formula represents the order of the respective variation.

$$\boldsymbol{f(i, d, AV_N) = \lfloor (i/K_d) \rfloor \,\%\, G_{Nd}} \quad (7)$$

The "K_d" variable utilized in the above formula and how this variable is calculated are not yet disclosed. This variable indicates in how many steps the genotypes of the d^{th} variation vary, that is, the step count of the d^{th} variation. For instance, if we look at Table 3, the length of steps in the 0^{th} variation is 3, that is, the genotypes change in every three steps. On the other hand, this value is 1 for the first variation. Equation (8) developed to compute the "K_d" value is given below. There are two different parts in this formula. In the upper section, the step count of the variation with the greatest level is determined. As can be seen from the formula, this value is always 1. In the bottom section, the step counts of other variations are calculated. Here, in order to find the step count of the variation at level d, the genotype numbers of the variations in the upper levels are multiplied, starting from the $(d+1)^{th}$ level.

$$\boldsymbol{K_d = \begin{cases} 1\,, & d == AV_N - 1 \\ \prod_{k=d+1}^{AV_N - 1} G_{Nk}\,, & d < AV_N - 1 \end{cases}} \quad (8)$$

Equation (7), as stated above, is used to find the genotype (numerical equivalent) of a given class for only one variation. If we want to calculate all the genotypes corresponding to a given class, then we must use (7) as the number of variations. The algorithm that accomplishes this process is given below. The "Convert_to_Genotypes()" algorithm has two parameters and these are the respective symbols of (7). Accordingly, the parameter "N_V" represents the number of variations, which is denoted with "AV_N" in (7). The line 3 of the algorithm corresponds to (7). Via the for loop in line 2, all the genotypes (numerical equivalents) corresponding to a given class are obtained uncomplicatedly by executing (7) as the number of variations. In the following algorithm, if we want to obtain genotypes instead of numerical equivalents, we need to convert the type of the "genotypes" variable into a two-dimensional string array and call the "Convert_into_genotype(index)" method on line 4 for "g".

```
Convert_to_Genotypes(int class_no, int N_V)
 1 genotypes is an integer array of length N_V
 2 for (int d = 0; d < N_V; d++)
 3     int g = (Math.Floor(class_no / Kd) % GNd)
 4     genotypes[d] = g;
 5 return genotypes;
```

We have stated before that the relation between the classes and the genotypes is bidirectional. Also, in the above section, we have shown the formulas and algorithms required to determine the genotypes of any class whose number is specified. The formulas and algorithms developed to create the second direction of the relation between classes and genotypes

will be described in this section. Equation (9) was developed to calculate the number of any class whose genotypes are given, in other words, to find the class to which any person belongs. Based on the genotypes of an individual, his/her class can be computed readily by using (9).

$$i = \sum_{d=0}^{AV_N - 1} G_{CCd} \times K_d \quad (9)$$

The algorithm Convert_to_Class_Number(), which corresponds to (9), is below. This algorithm takes only one parameter: a string array that holds the genotypes of the individual. In the fourth line of the algorithm, the numerical equivalent of the d^{th} genotype is calculated. On the other hand, the step count of the d^{th} variation is determined in the fifth line of the algorithm. Note that the calculation of the respective values ("G_{CCd}" and "K_d") in these two lines is not explicitly stated.

```
Convert_to_Class_Number(string[] genotypes)
1 int number1 = 0, number2 = 0;
2 int class_number = 0;
3 for (int d = 0; d < genotypes.Length; d++)
4     number1 = Compute GCCd
5     number2 = Compute Kd
6     class_number += (number1 * number2);
7 return class_number;
```

Individuals, who are genetically (disease-based) most similar to a given individual, are those that are involved in the same class with that given individual. Accordingly, the first place that must be looked is the class that individual belongs to. If there are no other individuals who are involved in the class of a given individual, other classes should be searched. At this point, experts can identify genetically most similar people to a person by looking at the relevant tables that are the outputs of the developed program.

B. Common Format for Relative Risk Models

One of the ordinary works done by the physicians working in the department of medical genetics is calculating the disease risks of individuals. In addition, clinicians who work in the departments such as heart, internal medicine and oncology search previous patients who are genetically most similar to a new patient and apply a successful treatment method previously applied. But, unfortunately, clinicians perform such tasks manually. This situation both leads to a waste of time and may cause confusion. Reaching all the relative risk models through a single application will be a significant convenience for clinicians/physicians. For that, before the development of the program, a common format was created for relative risk models. According to this approach, relative risk models are the inputs of the program and are independent from the program. Alterations made in any relative risk model or developing a novel relative risk model will not affect the program in any way. Thus, clinicians will be able to choose the relative risk model of any disease in an easy way through the program.

The allelic architecture (number, effect size, reference base, alternate base and frequency of susceptibility variants) differs across diseases. For instance, the number of identified variants is 5 for age-related macular degeneration (AMD), which is a common disease, whereas the number of identified variants is 32 for Crohn's disease. Apart from this, most variants identified confer increments at risk, whereas the remaining ones confer decrements at risk. Also, scientific research groups can sometimes handle the rare variants, which increase or decrease the disease risk at similar rates in a single category, in order to simplify the relative risk models as much as possible. During the development of the common format for the relative risk models, all these conditions were considered. An example relative risk model for age-related macular degeneration disease, which is created in accordance with the common format, is seen in Table 4. Although this common format is a powerful format, alternatively, a simpler format (e.g. the ids of the variants, the alleles of the variants and the allele's risk of developing the disease) can be produced and integrated into the program. However, in this case, there will be no need for the relative risk table, the third component of our program. The reason of this is that the relative risk table and the classification table will have the same number of elements. Therefore, the relative risk table should be disabled if the simple format mentioned is used.

TABLE IV. RELATIVE RISK MODEL FOR AMD

Number	**Genotypes (Four Fields)**				**Effect Size (Odds Ratio)**
	rs1061170	*rs1410996*	*rs10490924*	*rs9332739 or rs641153*	
	Type = Genotype	*Type = Number of risk alleles present*	*Type = Genotype*	*Type = Rare allele present or not*	
	Alleles = T,C	*Alleles = A,G*	*Alleles = G,T*	*Alleles = G,C&G,A*	
0	C|C	2	G|G	Yes	16.2
1	C|C	2	G|G	No	30.0
2	C|C	2	G|T	Yes	49.8
3	C|C	2	G|T	No	92.5
⋮	⋮	⋮	⋮	⋮	⋮

The upper side of the thick line in Table 4 is called header part, whereas the underside is called values part. In the header part, the ids of the variants or variants group, which variants are handled together, the type of the variant or variant group in the risk model ("Genotype", "Number of risk alleles present", "Rare allele present or not") and the alleles of the variants must be specified. In the values part, genotype combinations and each genotype combination's risk of developing the disease (odds ratio) are contained. Any relative risk model constructed according to the common format can be stored as excel spreadsheet or can be stored in database.

C. The Dynamic Program and Its Components

The features such as computing the disease risks of a large number of people simultaneously and classifying people according to their genetic characteristics, are very important for personalized medicine and preventive medicine. Therefore, considering these features, a dynamic program was developed which is compatible with the relative risk models (formed in accordance with the common format).

The developed program has basically two inputs. The first of them, as we have mentioned above, is the relative risk model of any disease, formed in accordance with the common format.

The second input is individuals' genotypes, regarding the variations associated with the disease. These personal genotypes, which are the second input of the program, are brought dynamically from the database. On the other hand, in systems / clinics where personal genotypes are organized as excel spreadsheets, the second input of the program may also be an excel spreadsheet. An example table for individuals and their genotypes, regarding age-related macular degeneration disease, is seen in Table 5. By taking these two inputs, our program classifies individuals based on their genotypes and computes the disease risks of them.

Our program is based on the basic classification approach, which is described in the upper section, and consists of four main components. These components are "Variant Array", "Classification Table", "Relative Risk Table" (different from the relative risk model), and "Hash-Table". Classification table and relative risk table are shown in Fig. 1 and in Table 7, respectively.

TABLE V. AN EXAMPLE TABLE OF INDIVIDUALS AND THEIR GENOTYPES

Individuals	**Genotype Values**				
	rs1061170	*rs1410996*	*rs10490924*	*rs9332739*	*rs641153*
Individual4	T|C	G|G	G|T	G|G	G|A
Individual9	C|T	G|G	G|T	G|G	G|G
Individual47	T|T	A|A	G|G	G|G	G|G
⋮	⋮	⋮	⋮	⋮	⋮

1) Variant Array: The first and the simplest one from the components constituting our program is "Variant Array". A sample variant array is seen in Table 6. Variant array, as is evident from its name, stores the information concerning the variations specified in the relative risk model. Also, each element of the variant array consists of two attributes. The first attribute holds the id of the variation. On the other hand, the second attribute is "Alleles" array and, this attribute is used in some formulas in the basic classification approach.

TABLE VI. A SAMPLE VARIANT ARRAY

Indices	**Values**
0	rs1061170, (T, C)
1	rs1410996, (A, G)
2	rs10490924, (G, T)
⋮	⋮

After the program reads the relative risk model, firstly, it stores the variations defined in the header portion of the relative risk model into the variant array. Both the order of the variations and the order of the alleles in the header portion are preserved during the recording process. If there are variations addressed together in the header portion, program stores them one by one into the variant array after decomposing. After the storage of whole variations present in the header portion into the variant array, variation-related operations are now performed only through the variant array and we do not have to read the relative risk model again and again.

2) Classification Table: The second component constituting our program is "Classification Table" and this table is basically used to store individuals (as classified based on their genotypes). A sample classification table is seen in Fig. 1. Classification table is actually an array and the indices of the array are the class numbers. The size of the array is found according to (2). The parameter "AV_N" in (2), namely, the number of variations associated with the disease, is the size of the variant array. Therefore, after all the variations specified in the header portion of the relative risk model are stored into the variant array, classification table is created dynamically during the execution of the program. Each element of the classification table comprises of an integer variable showing which group of the relative risk table corresponds to that class, a linked-list storing individuals who are in that class and who have the same genotype and, lastly, a string array storing the genotype combinations in open format. (Note that the third component, string array, is not shown in Fig. 1.)

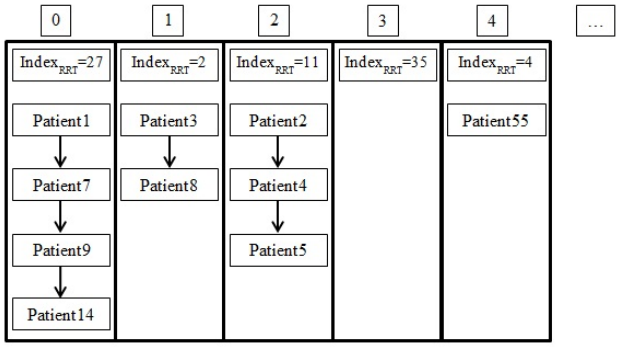

Fig. 1. A sample classification table

After the personal genotypes, which are the second input of the program, are fetched from the database, the Convert_to_Class_Number() algorithm, corresponding to (9), is executed for each individual and the classes of the individuals are determined. In addition, individuals are added to the list of classes, which they belong to. Other characteristics of the classification table are as follows: Array indices and integer variables (group number – index of relative risk table) establish a connection between classification table and relative risk table. On the other hand, by default, "-1" are assigned to the integer variables during the creation of the classification table. A value of -1 indicates that the class does not correspond to any group in the relative risk table. When relative risk table is being created, the values of the integer variables in the classification table are updated at the same time.

TABLE VII. A SAMPLE RELATIVE RISK TABLE SORTED BASED ON THE EFFECT SIZES

Indices (Group No)	**Values**		
	Effect Size (Odds Ratio)	*Genotype Preferences*	*List of Corresponding Classes*
0	1.0	(T|T, 0, G|G, Y)	1->2->3->4->5->6->7->8
1	1.9	(T|T, 0, G|G, N)	0
2	2.7	(T|T, 2, G|G, Y)	55->56->57->58->59->60->61->62
3	2.7	(T|T, 1, G|G, Y)	28->29->30->31->32->33->34->35
⋮	⋮	⋮	⋮

As mentioned before, in the case that a simpler format (e.g. the ids of the variants, the alleles of the variants and the allele's risk of developing the disease) is used for relative risk models, there will be no need for the relative risk table. Therefore, the relative risk table should be disabled and the elements (classes) of the classification table should be sorted based on the risk values if the simple format mentioned is used.

3) Relative Risk Table: The third component of our program is "Relative Risk Table". A sample relative risk table is seen in Table 7. Although relative risk table is very similar to the relative risk model, these two structures are distinct from each other. Relative risk models are arduous models that are generated by scientific research groups doing genome-wide association studies. These models show the disease-related variations and at what rate the genotypes influence the disease. On the other hand, relative risk table is a data structure that we developed, and it is a component of our program. Any relative risk model formed in compliance with the common format is converted into the corresponding relative risk table by our program. The relative risk table is also an array and the indices of the array are the group numbers. The size of this array is the same as the size of the relative risk model. Each element of the relative risk table consists of a float variable indicating the disease risk (Odds Ratio) of that group, an integer list storing which class/classes in the classification table that group corresponds to and a string array storing the combinations of the genotypes of that group.

4) Individual Hash Table: The fourth and the last component of our program is a hash-table, which stores individuals and classes they belong to. A sample hash-table is seen in Table 8. While explaining the classification table, we stated that the person's class is determined, and the person is added to the list of that class after the personal genotypes are fetched from the database. A similar process also applies to the hash table. Personal genotypes are brought from the database together with person information. In parallel, after the reading process, individuals, and classes they belong to are stored also in the hash table.

TABLE VIII. A SAMPLE HASH TABLE HOLDING INDIVIDUALS AND THEIR CLASSES

Indices	Values
0	(Individual1, 4)
1	
2	(Individual4, 13)
3	(Individual3, 2)
4	
5	(Individual9, 0)
⋮	⋮

We established a relation between the hash table and the classification table by storing individuals' classes in the hash table. We mentioned that there is a similar connection between classification table and relative risk table. Accordingly, both the connection between the hash table and the classification table and the connection between the classification table and the relative risk table are used to obtain the disease risk of any person. To find out the result of such a query, firstly, the person is searched in the hash-table and the class of the person is obtained, secondly, the relative risk table group corresponding to the person's class is determined from the classification table, and lastly, the disease risk of that group is obtained from the relative risk table. Since the person belongs to that group in the relative risk table, the risk of the person is the risk of that group. As depicted above, both the connections between the components and the hash-table significantly reduce the cost of search operations.

III. RESULTS AND DISCUSSION

The developed method and program were applied on the variation-based personal genetic data published by 1000 Genomes Project and the relative risk model generated for the age-related macular degeneration disease. The relative risk model that was used is seen in Table 4. After loading the data to the program, 2504 people were classified according to age-related macular degeneration disease. The classification table produced by the program is seen in Table 9 (Only a certain part of the classification table is shown here.).

TABLE IX. CLASSIFICATION TABLE AFTER AMD DISEASE RISK MODEL APPLIED ON THE SAMPLES OF 1000 GENOMES PROJECT

Class No	Genotype Preferences	Number of Individuals	Relative Risk No (Group No)	Risk
0	(T|T, A|A, G|G, G|G, G|G)	240	1	1.9
1	(T|T, A|A, G|G, G|G, G|A)	56	0	1
2	(T|T, A|A, G|G, G|G, A|A)	8	0	1
⋮		⋮	⋮	⋮
62	(T|T, G|G, G|G, C|C, A|A)	0	2	2.7
63	(T|T, G|G, G|T, G|G, G|G)	68	15	15.3
64	(T|T, G|G, G|T, G|G, G|A)	10	10	8.3
⋮		⋮	⋮	⋮
188	(C|C, A|A, T|T, C|C, A|A)	0	-1	-
189	(C|C, A|G, G|G, G|G, G|G)	1	-1	-
190	(C|C, A|G, G|G, G|G, G|A)	3	-1	-
⋮		⋮	⋮	⋮
240	(C|C, G|G, T|T, C|C, G|G)	0	33	154
241	(C|C, G|G, T|T, C|C, G|A)	0	33	154
242	(C|C, G|G, T|T, C|C, A|A)	0	33	154

In Table 9, the class number, the number of individuals in that class, the corresponding group number (in the relative risk table) and the risk of getting the disease for the individuals in that class are shown side by side. Note that the risk information is not normally included in the classification table, but it is shown here only for convenience. Depending on the number of variations in the relative risk model and (2), there are 243 separate classes in the classification table. Here, one remarkable case is that the group numbers corresponding to some classes is -1, that is, these classes do not correspond to any group (in the relative risk table). The emergence of this situation is due to the relative risk model (scientific research group's desire of simplifying the model as much as possible). For instance, although there are 243 different classes for this disease, there are only 36 groups in the relative risk model. Namely, the research group which created the model reduced

243 possible genotype combinations to 36 groups (depending on the allele frequencies in European population).

The relative risk table generated by the classification program is shown in Table 10. There are 37 groups in the relative risk table, and the first 36 groups correspond to 36 groups in the relative risk model created for AMD disease. The last group of the table contains classes that do not correspond to any group, and the number of this group is specified as -1. In addition, the list of classes corresponding to the relevant group, the number of individuals in that group, and the risk of that group are also indicated in Table 10.

TABLE X. RELATIVE RISK TABLE AFTER AMD DISEASE RISK MODEL APPLIED ON THE SAMPLES OF 1000 GENOMES PROJECT

Group No	**List of Classes**	**Genotype Preferences**	**Number of Total Individuals**	**Risk**
0	1->2->3->4->5->6->7->8	(T|T, 0, G|G, Y)	74	1
1	0	(T|T, 0, G|G, N)	240	1.9
2	55->56->57->58->59->60->61->62	(T|T, 2, G|G, Y)	19	2.7
⋮	⋮		⋮	⋮
33	235->236->237->238->239->240->241->242	(C|C, 2, T|T, Y)	4	154
34	153	(C|T, 2, T|T, N)	9	190
35	234	(C|C, 2, T|T, N)	6	285
-1	81->82->83->84->85->86-> . . .	-	20	-

The proposed program classifies individuals according to their genetic characteristics (the genotypes of the individual for the variations indicated in the relative risk model of the disease concerned) and produces the relevant outputs (tables). Finding the people who are genetically most similar to any person is a process that should be performed by the experts who use the program, not the program. This is because, for diagnostic or therapeutic purposes, different parameters and information may need to be considered in the process of finding the people who are genetically most similar to any person, and it is the specialist personnel who need to do this. On the other hand, even if there is no different parameter or information, experts should use the tables produced by the program and the relations between tables effectively in this process. From these perspectives, the decision maker is the expert staff, and the program only systematically presents the relevant information to the decision maker.

After the classification process is complete and the tables are generated, if the doctor wants to find the genetically most similar individuals to any individual, he or she should follow these steps: First, the person must be searched. When the person is found, the program will be positioned to the row in the hash table, where the person is located. At the same time, the program will be positioned to the row in the classification table, where the respective class is located. From the classification table, all persons in that class can be accessed, and besides, the genotype characteristics of that class, which group the class corresponds to in the relative risk table, and the relative risk value of the class can be obtained. The genotypes of all individuals in a class (regarding the variations indicated in the relative risk model) are exactly the same, that is, the individuals genetically most similar to each other are those in the same class. If there is no one else in the respective class, or if those present in the class do not provide adequate information about diagnosis or therapy, then the doctor will need to search in other classes. At this point, there are three options: 1) Other classes (if any) in the same group as the relevant class can be looked at in the relative risk table. 2) By making changes on the genotype characteristics of the relevant class, it can be easily found which class the new genotype combination corresponds to. 3) All other classes that have a different combination (1 allele different, 2 alleles different, and so on) from the genotype combination of the respective class can be easily found. The operations specified in the second and third options can be easily performed through the relevant parts of the program's interface. Thanks to the formulas and algorithms running in the background, the program returns the desired results simply and quickly. At this point, the only thing that the doctor must do is enter/select the values and press the button. As an example for the second option, assume that the doctor changed 2 alleles on the genotype combination (T|T, A|A, G|G, G|G, A|A) of class 2 and formed a new genotype combination as (C|C, A|A, G|G, G|G, A|A). If the doctor presses the corresponding button after entering the new genotype combination, the program will return that this new genotype combination corresponds to class 164. Besides, the program will be positioned to the row in the classification table, where class 164 is located. From the classification table, all individuals in class 164 can be accessed.

IV. CONCLUSION

In this study, a novel dynamic method and program is proposed for the disease-based genetic classification of individuals. Our generic classification method and program can classify individuals according to their disease-based genetic characteristics and can calculate disease risks of them simultaneously. The basic classification approach is completely based on the mathematical formulas and supports all types of variations. On the other hand, the common format was designed for the relative risk models, considering the common preferences of them. Our common format does not support relative risk models that include factors other than genetic factors, such as age, gender, smoking, etc. Our common format only supports relative risk models based on genetic factors. In parallel, the dynamic application, which is constructed on the basic classification approach, can work properly with the relative risk models developed for different diseases. Through this program, relative risk models can be managed from a single point, many people can be classified based on their genetic characteristics and the disease-risks can be calculated. On the other hand, calculating genetic similarities between classes is beyond the scope of this study. Depending on that, on the basis of disease, people who are genetically most similar to a person can be identified by only experts, using the outputs of the program (related tables). In short, this study contributes to personalized medicine approaches to some extent.

The relative risk model generated for the age-related macular degeneration disease and the personal variation data of 2504 people published by 1000 genomes project were applied to the program presented in this paper. With the loading of the relevant data to the program, 2504 people were classified according to age-related macular degeneration disease, the relative risks of these individuals were calculated, and the relevant tables, which are the outputs of the program, were produced. With the same logic, by using the relative risk

model for any disease, which is constructed in accordance with the common format, a large number of individuals can be classified according to the respective disease.

REFERENCES

[1] J. Hardy, A. Singleton, "Genomewide association studies and human disease", *New England Journal of Medicine*, *360*(17), 1759–1768, 2009.

[2] J. Krier, R. Barfield, R.C. Green, P. Kraft, "Reclassification of genetic-based risk predictions as GWAS data accumulate", *Genome medicine*, *8*(1), 1-11, 2016.

[3] Internet: GWAS Catalog, https://www.ebi.ac.uk/gwas/, 30.07.2021.

[4] T. A. Manolio, F. S. Collins, N. J. Cox, et. al., "Finding the missing heritability of complex diseases", *Nature*, *461*, 747–753, 2009.

[5] T. Beck, T. Rowlands, T. Shorter, A. J. Brookes, GWAS Central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies, Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023, Pages D986–D993, https://doi.org/10.1093/nar/gkac1017.

[6] Hettiarachchi, G., & Komar, A. A. (2022). Genome Wide Association Studies (GWAS) to Identify SNPs Associated with Common Diseases and Individual Risk. In Single Nucleotide Polymorphisms: Human Variation and a Coming Revolution in Biology and Medicine (pp. 51-76). Cham: Springer International Publishing.

[7] L. A. Hindorff, P. Sethupathy, H. A. Junkins, et. al., "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits", *Proceedings of the National Academy of Sciences*, *106*(23), 9362-9367, 2009.

[8] S. J. Schrodi, S. Mukherjee, Y. Shan, et. al., "Genetic-based prediction of disease traits: Prediction is very difficult, especially about the future", *Frontiers in genetics*, *5*, 162, 2014.

[9] Lee, M. J., Lee, I., & Wang, K. (2022). Recent advances in RNA therapy and its carriers to treat the single-gene neurological disorders. Biomedicines, 10(1), 158.

[10] M. M. Alves, Y. Sribudiani, R. W. W. Brouwer, et. al., "Contribution of rare and common variants determine complex diseases-Hirschsprung disease as a model", *Developmental biology*, *382*(1), 320-329, 2013.

[11] J. Altmüller, L. J. Palmer, G. Fischer, et. al., "Genomewide scans of complex human diseases: True linkage is hard to find", *The American Journal of Human Genetics*, *69*(5), 936-950, 2001.

[12] J. C. Barrett, S. Hansoul, D. L. Nicolae, et. al., "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease", *Nature genetics*, *40*(8), 955-962, 2008.

[13] J. Maller, S. George, S. Purcell, et. al., "Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration", *Nature genetics*, *38*(9), 1055-1059, 2006.

[14] E. Zeggini, L. J. Scott, R. Saxena, et. al., "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes", *Nature genetics*, *40*(5), 638-645, 2008.

[15] K. Yasuda, K. Miyake, Y. Horikawa, et. al., "Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus", *Nature genetics*, *40*(9), 1092-1097, 2008.

[16] S. Kathiresan, B. F. Voight, S. Purcell, et. al., "Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants", *Nature genetics*, *41*(3), 334, 2009.

[17] Weeks, Elle M., et al. "Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases." Nature Genetics 55.8 (2023): 1267-1276.

[18] Abdellaoui, A., Dolan, C. V., Verweij, K. J., & Nivard, M. G. (2022). Gene–environment correlations across geographic regions affect genome-wide association studies. Nature genetics, 54(9), 1345-1354.

[19] C. Sabatti, S. K. Service, A. L. Hartikainen, et. al., "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population", *Nature genetics*, *41*(1), 35-46, 2009.

[20] W. Zheng, J. Long, Y. T. Gao, et. al., "Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1", *Nature genetics*, *41*(3), 324-328, 2009.

[21] C. Katsios, D. H. Roukos, "Individual genomes and personalized medicine: Life diversity and complexity", *Personalized Medicine*, *7*(4), 347-350, 2010.

[22] M. A. Hamburg, F. S. Collins, "The path to personalized medicine", *New England Journal of Medicine*, *363*(4), 301-304, 2010.

[23] G. S. Ginsburg, J. J. McCarthy, "Personalized medicine: Revolutionizing drug discovery and patient care", *TRENDS in Biotechnology*, *19*(12), 491-496, 2001.

[24] N. J. Schork, "Personalized medicine: Time for one-person trials", *Nature News*, *520*(7549), 609, 2015.

[25] Yamamoto, Y., Kanayama, N., Nakayama, Y., & Matsushima, N. (2022). Current status, issues and future prospects of personalized medicine for each disease. Journal of Personalized Medicine, 12(3), 444.

[26] Hassan, M., et. al. (2022). Innovations in genomics and big data analytics for personalized medicine and health care: A review. International journal of molecular Sciences, 23(9), 4645.

[27] The International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome", *Nature*, *431*(7011), 931-945, 2004.

[28] S. Levy, G. Sutton, P. C. Ng, et. al., "The diploid genome sequence of an individual human", *PLoS biology*, *5*(10), 2113–2144, 2007.

[29] International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs", *Nature*, *449*(7164), 851, 2007.

[30] 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing", *Nature*, *467*(7319), 1061–1073, 2010.

[31] 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes", *Nature*, *491*(7422), 56-65, 2012.

[32] 1000 Genomes Project Consortium, "A global reference for human genetic variation", *Nature*, *526*(7571), 68-74, 2015.

[33] 1000 Genomes Project Consortium, "An integrated map of structural variation in 2,504 human genomes", *Nature*, *526*(7571), 75-81, 2015.

[34] Internet: 1000 Genomes Project Consortium, /vol1/ftp/release/20130502/ directory, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/, 30.07.2021.

[35] Internet: 1000 Genomes Project Consortium, /vol1/ftp/release/20130502/supporting/bcf_files directory, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/bcf_files, 30.07.2021.