Enabling the Use of Hereditary Information from Pedigree Tools in Medical Knowledge-based Systems

Pablo Gay, Beatriz López, Albert Plà, Jordi Saperas, Carles Pous

University of Girona, Campus Montilivi, P4 Building,
Girona, E17071, Spain,
{pablo.gay, beatriz.lopez, albert.pla, jordi.saperas, carles.pous}@udg.edu

Abstract

The use of family information is a key issue to deal with inheritance illnesses. This kind of information use to come in the form of pedigree files, which contain structured information as tree or graphs, which explains the family relationships. Knowledge-based system should incorporate the information gathered by pedigree tools to assess medical decision making. In this paper, we propose a method to achieve such a goal, which consists on the definition of new indicators, and methods and rules to compute them from family trees. The method is illustrated with several case studies. We provide information about its implementation and integration on a case-based reasoning tool. The method has bee experimentally tested with breast cancer diagnosis data. The results show the feasibility of our methodology.

Keywords: Decision support, Structured data, Feature construction, Modeling, Medical applications

1. Introduction

Family history has been important for preventing several inheritance diseases, as it is one of the key variables in the Gail model [1] for breast cancer diagnosis. Family information is usually gathered thanks to pedigree software (as Cyrillic [2]), which allows to annotate relationships, healthy states, genetic markers, and much more data on patients and relatives, in a tree structured way. Thus, to improve medical care, knowledge based system should incorporate the information about families collected thanks to this kind of tools.

Data processing on pedigree software has been traditionally faced from a statistical point of view. However, statistics are not easy in this kind of structured scenarios, most popular statistics tools can conduct to inappropriate or absurd conclusions, and other methods for compositional data are required. On the other hand, expert physicians can evaluate at a glance, from the structure, density and another heuristic knowledge, the risk of a member of the family for suffering an inheritance illness. The skill of evaluating the information of the family is something that is acquired by experience, and difficult to transmit to other, novice physicians. Our research concerns the development of tools that capture the heuristic knowledge of expert physicians, finding out measures from the tree structure that conducts as close as possible to the predictions made by them. Providing a method to extract the relevant information from family trees enables the integration of pedigree tools with medical knowledge based system so other physicians can also use

inheritance data in their decision making.

The contribution of this paper is our methodology towards achieving such integration. It includes the definition of structured data-based indicators which are computed by analyzing the information contained in pedigree files. The methodology is presented first under the assumption of a simple, hierarchical family, and then is extended to cover more complex situations (second marriages, and so on). Our research is constrained to the data we have on breast cancer, an illness in which inheritance has been proved to be a key factor. Nevertheless, we believe that other inheritance illnesses can benefit from our results.

This paper is organized as follows. First we provide information about the structured data on Section 2. Next, in Section 3, we describe our methodology to evaluate a set of indicators from pedigree files. In Section 4 case studies are provided, and in Section 5 the experimentation performed so far is shown and discussed. Then, in Section 6 we expose some related work and, finally, we end the paper in Section 7 with some conclusions and future work.

2. Structured Family Data

Our starting point is the family information gathered in the very well-known standard that nowadays is one of the most used for pedigree information sharing: the GEDCOM format (GEnealogy Data COMmunication) [3]. This format consists of a header section, records, and a trailer section. Within

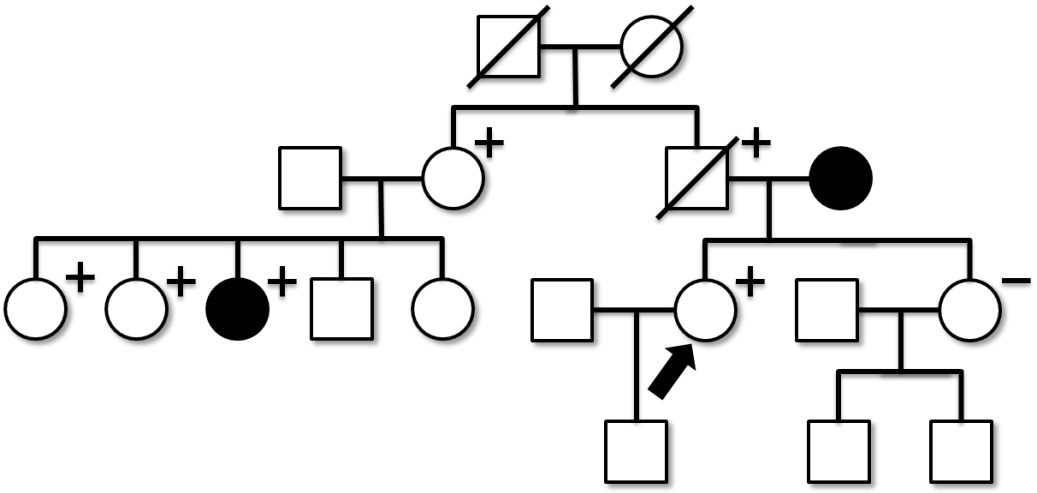

Figure 1: This figure represents an example of a GEDCOM file pedigree tree created using Cylillic. Family members are represented through squares (males) and circles (females) and relationships using lines.

these sections, records represent people, families, sources of information, and other miscellaneous records, including notes. In our case, the information we require is the family relationships and the people's relevant medical data records such as if she is affected by a disease or has some results on a previously performed genetic test.

As shown in Figure 1, the representation of a GEDCOME file allows physicians to quickly understand the family structure and also the inherited factors on the members. Squares/circles represent males/females members of the family respectively; members in the same horizontal level belong to the same generation, horizontal lines between members in the same generation represents marriage relations, and vertical lines between members who belong to different generations represent parenthood and childhood relations.

Regarding the individual information, a crossed line over the member means she is deceased; when the member is in black it means that she is affected by the illness, a plus next to the member means that the member has been genetically tested and has the disease causing mutation, a minus means the member has no the mutation and nothing implies that he has not been tested correspondingly.

3. Methodology

With the information included in the pedigrees, we can extract information in the form of indicators and use them into a knowledge-based system to estimate the risk of suffering the illness. Several indicators can be defined, depending if we want to evaluate the family as a whole or at the individual level. In the former case, statistics-based indicators can be used, while in the second case, the value of an indicator assigned to an individual depends on its position on the family tree. Then, the structure of the family tree is important, and new, structure data-based methods are required. Figure 2 shows the different indicators presented in this paper. They can be combined or not depending on the particularities of the medical application. All of the indicators can feed a medical knowledge-based system to support medical decision making.

Regarding the interpretation of the family structure, an extension to the method is required to appropriately compute the indicators in complex pedigrees with multiple roots. Such extension is presented at the end of this

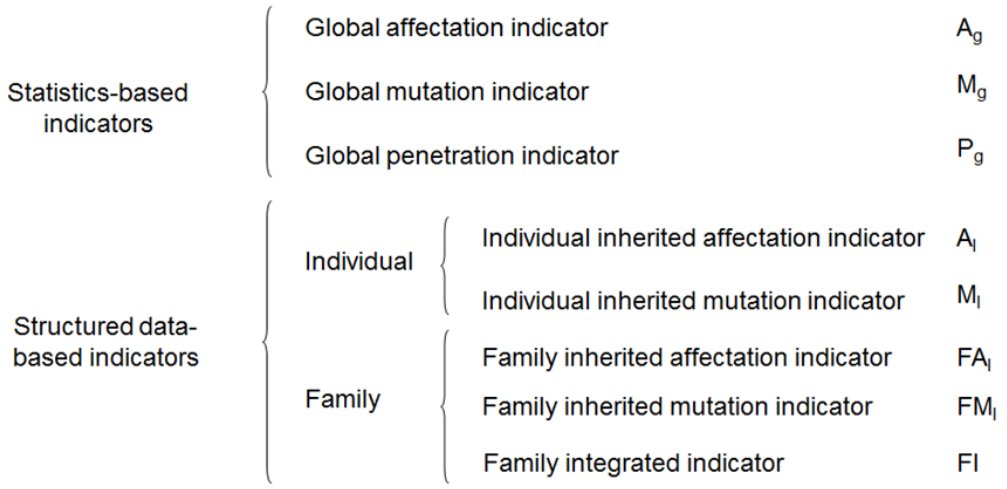

Figure 2: Indicators for inheritance illnesses risk assessment.

section.

3.1. Statistic-Based Indicators

Statistics-based indicators are the ones currently used by physicians and provide general information about the pedigree. They can be estimated without having any knowledge on the pedigree structure. We have considered three of them: the global affectation, the global mutation, and the global penetration indicator.

3.1.1. Global Affectation Indicator

The global affectation indicator is one of the most basic statistic-based indicators, because it represents the probability of affected family members regarding the whole population (in our case, the family). The global affectation indicator A_g is formally defined as follows:

$$A_g = \frac{A}{T} \tag{1}$$

Where:

- A stands for the total amount of people who had or have suffered the illness.
- T stands for the total amount of family members in the pedigree (family) under study.

For example, suppose the pedigree shown at Figure 3 composed by 10 members, two of them having developed the illness (members 4 and 7). Therefore, the global affectation indicator is $A_g = 2/10 = 0.2$).

3.1.2. Global Mutation Indicator

The global mutation indicator estimates the probability of being a carrier of the mutated gene responsible of the disease, regarding the population. The global mutation indicator, M_g is defined as follow:

$$M_g = \frac{M}{T} \tag{2}$$

Where:

- M stands for the total of people who had or have mutated genes responsible of the inheritance disease.

Following the example of Figure 3, in this case there are four family members who have been tested positive for the genetic predisposition (2, 4, 6 and 8), so the global mutation indicator is $M_g = 4/10 = 0.4$).

3.1.3. Global Penetration Indicator

The global penetration indicator represents how aggressive is the specific mutation which affects the pedigree. Specifically, the global penetration indicator P_g informs about how many of the mutations have actually become an affectation. It is computed as follows:

$$P_g = \frac{A}{M} \tag{3}$$

With this indicator, we can know how probable is that a carrier becomes an affected. Again, in the example of Figure 3, there are four members who carry the responsible gene (members 2, 4, 6 and 8) but actually just two of them developed the illness (members 4 and 7), hence the global penetration indicator is $P_g = 2/4 = 0.5$.

Statistics-based indicators can be few discriminative, since they provide the same information to all of the members of the family independently of the branch of the family. Our proposal is to complement it with the structured data-based indicators.

3.2. Structured Data-Based Indicators

Structured data-based indicators include information about the pedigree structure. This kind of indicators allows differentiating between members

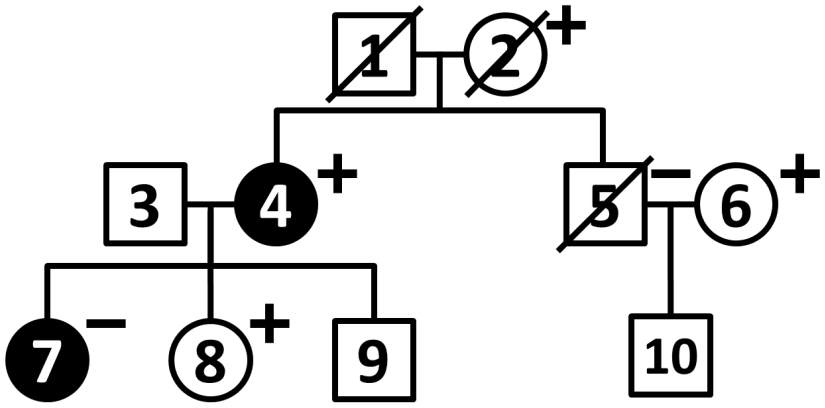

Figure 3: Family tree simple example. The statistics-based indicators are $A_g = 2/10 = 0.2$, $M_g = 4/10 = 0.4$ and $P_g = 2/4 = 0.5$.

of two pedigrees with the same amount of affected and mutated members by considering where these mutations and affectations are and how are they related to the other members. The indicators are defined at individual and family levels, both, for affectation and mutation information. The former analyze the information concerning a single member of the family, while the latter aggregates the individual indicators for a given family.

3.2.1. Individual Inherited Indicator

This indicator defines the ratio regarding the amount of family members that are or were affected by the illness taking into account their relationships.

To estimate the inherited affectation indicator of a member of a pedigree we need two information pieces: the ancestor's history and the indirect diversification factor. First, the ancestor's history of a member $\delta(m)$ is based on the generation era: the oldest family member, the one at the top of the

family tree, is assigned the era zero, his sons the era one, his grandsons the era two and so on. Then, $\delta(m)$ estimates the history of the healthy/affected ancestors of a node, according to the following equation:

$$\delta(m) = \sum_{k=1}^{era(m)} state_k \times 2^k \tag{4}$$

Where:

- era(m) is the generation era of the m member.
- $state_k$ is 0 if the k ancestor of m is healthy; 1 otherwise

Marriages are managed as a unity; the healthy state of marriage node is computed as the worst case of the couple.

To illustrate this indicator, let us suppose the situation of Figure 4. The era values are shown at the left of the Figure.

Members A to F are leaf nodes, while $N1$ to $N5$ are inner nodes with offspring. The ancestors of node A are: $N4$ (era 2), $N2$ (era 1), $N1$ (era 0). Since there is a single affected ancestor in the history of A, $N1$, the ancestor history of A is finally computed as $\delta(A) = (1 \times 2^0) + (0 \times 2^1) + (0 \times 2^2) = 1$. Analogously, nodes B, D, and F have $\delta(B) = \delta(D) = \delta(F) = 1$, and the inner nodes $\delta(N4) = \delta(N2) = 1$. Leaf node C is also affected so its value is $\delta(C) = 2^0 + 2^2 = 5$. Node E is affected as well as its two first ancestors, therefore $\delta(E) = 2^0 + 2^1 + 2^3 = 11$. Finally, node F is the same case as node E regarding their common ancestors but it is not affected, so the value is $\delta(F) = 2^0 + 2^1 = 3$.

Algorithm 1 : A_I(node, era, idf, delta)

1: **if** node is affected **then**
2: $newDelta = delta + 2^{era}$
3: **else**
4: $newDelta = delta$
5: **end if**
6: **if** node has offspring **then**
7: $indicator = 0$
8: $newSegment = \frac{idf}{offspring\ size}$
9: $newEra = era + 1$
10: **for all** p in node's offspring **do**
11: $indicator\ +\!= A_I(p, newEra, newSegment, newDelta)$
12: **end for**
13: **return** $indicator$
14: **else**
15: **return** $newDelta \times idf$
16: **end if**

Second, the indirect diversification factor represents how many times the precedence branch of a node has been split. The more a branch splits, the less value it has. To compute it, we assign a unitary length segment to the top node representing the tree root. Then the segment is split into equal sub-segments, one for each descendant node. The descendant nodes repeat this process (but just with the sub-segment they were assigned, not the whole 1-length segment) until the node under study is reached. The indirect diversification factor of a node, IDF(m), is computed as the length of his corresponding sub-segment.

To illustrate the IDF, an example is provided in Figure 4. The top node ($N1$) has been assigned a 1-length line segment. Since it has two descendant nodes, those have a 1/2-length segment each one. Then, on the left tree

branch (i.e., following down $N2$) there are three descendants, therefore the 1/2-length on the left is split into three 1/6-length segments; as C and D are leaves, then $IDF(C) = IDF(D) = 1/6$. Since the left node is still not a leaf, the 1/6-length is again split into two sub-segments attaching a 1/2-length line segment to nodes A and B (i.e. $IDF(A) = IDF(B) = 1/12$). The right tree branch proceeds in the same way, obtaining $IDF(E) = IDF(F) = 1/12$.

Finally, the inherited affectation indicator of a node $A_I(m)$ is computed as a function of its ancestors $\delta(m)$ and its IDF(m). Inspired in grounded mathematical models [4], we have chosen the times function to combine both components, as follows:

$$A_I(m) = IDF \times \delta(m) \tag{5}$$

Algorithm 1 summarizes the method to compute the individual inherited affectation indicator.

3.2.2. Individual Inherited Mutation Indicator

The inherited mutation indicator represents the probability with which a concrete family member could inherit a mutation by considering her ancestors information.

Given a member m of the family, the inherited mutation indicator, $M_I(m)$, is defined as follows:

$$M_I = \frac{|MA(m)|}{|Anc(m)|} \tag{6}$$

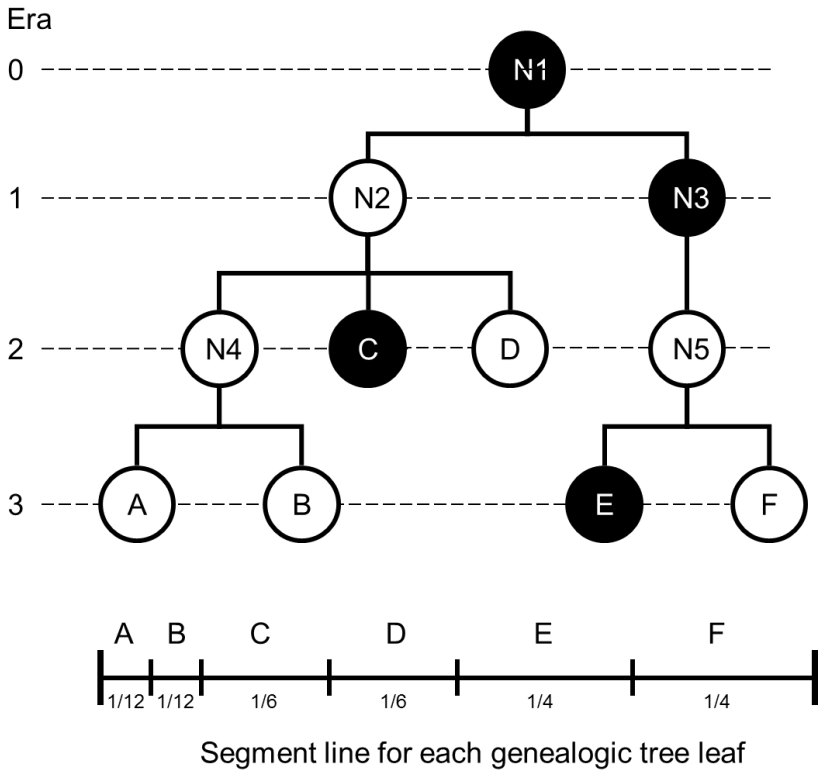

Figure 4: Example of the individual inherited affectation indicator.

Where:

- $Anc(m)$ is the set of valid ancestors of m (see below).
- $MA(m) \subseteq Anc(m)$ is the set containing all the family ancestors of m who have the mutated gene.

Physicians know how to explore appropriately the ancestors of a given member. After several interviews, we have acquired a set of rules to be applicable in the search of the ancestor of a given member. They are provided in Table 1. A graphic example for each rule is provided in Figure 5. Then,

the set of valid ancestors, $Anc(m)$, is computed according to these rules.

Table 1: Rules for tree exploration (bottom-up search).

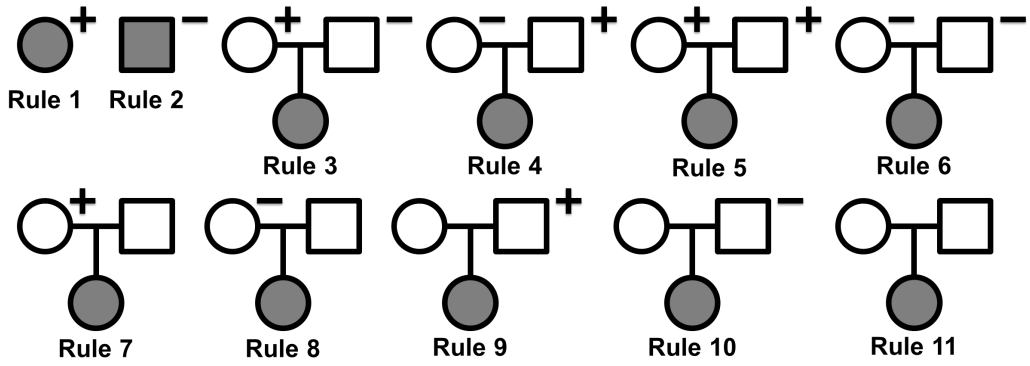

Figure 5: Graphic representation of the conditions of the ancestor's exploration rules. The gray node represents the family member under test.

For example, suppose we want know the M_I of the member 9 of Figure 3. The immediate ancestors of 9 are 3 and 4. They trigger rule 7 condition (node 4 is female with positive testing, and father is unknown); thus node 4 is selected according to the rule's consequent. Next, the ancestors of 4 are analyzed: 1 and 2. They trigger rule 7 again, and then the female 2 is selected. Since there is no more information about 2 ancestors, the bottom-up search is stopped. The set of valid ancestors carrying out information about the illness that can influence the member under study is $Anc(9) = \{4, 2\}$. Both ancestors have the mutation proved positive, so $M_A(9) = 4, 2$. Therefore, $M_I(9) = 2/2 = 1$. Note then the difference between the statistics-based indicator regarding mutation obtained in the previous section M_g and the new value computed for the individual member 9.

Rule	Conditions (Premise)	Action (Consequent)
Rule 1.	Actual family member tested positive for the genetic mutation.	No ancestor is explored, due we have the certainty that this member has the mutated gene.
Rule 2.	Actual family member tested negative for the genetic mutation.	No ancestor is explored, due we have the certainty that this member does not have the mutated gene.
Rule 3.	Mother tested positive and father negative for the genetic mutation.	Mother's ancestors are explored due the father has been tested negative.
Rule 4.	Mother tested negative and father positive for the genetic mutation.	Father's ancestors are explored due the mother has been tested negative.
Rule 5.	Both mother and father tested positive for the genetic mutation.	Both mother and father's ancestors are explored.
Rule 6.	Both mother and father tested negative for the genetic mutation.	There is no need to continue exploring because neither mother or father could propagate the mutated gene.
Rule 7.	Mother tested positive for the genetic mutation and father is unknown.	Mother's ancestors are explored.
Rule 8.	Mother tested negative for the genetic mutation and father is unknown.	Father's ancestors are explored due the mother has been tested negative.
Rule 9.	Father tested positive for the genetic mutation and mother is unknown.	Father's ancestors are explored.
Rule 10.	Father tested negative for the genetic mutation and mother is unknown.	Mother's ancestors are explored.
Rule 11.	Both mother and father are unknown.	Father's ancestors are explored if there are no mutated genes between mother's ancestors.

3.2.3. Family Inherited Indicators

Previous individual indicators can be applied to all of the youngest members of the family, representing the current generation, in order to obtain a family indicator, both, for the affectation and the mutation information. Moreover, they could be combined in a single indicator (family integration indicator) to assess about the risk of a family member to suffer the illness.

Family Inherited Affectation Indicator. The family inherited affectation indicator, FA_I, regards the information of the inherited affectation indicator of the youngest members of a family, i.e. of the leaves of a family tree. It is computed by aggregating the results of all the leaves, as follows:

$$FA_I = \sum_{\forall l | l \text{ is leaf}} IDF(l) \times \delta(l) \tag{7}$$

And following the example of Figure 4 we obtain the following results:

$$FA_I = ID_A \times \delta(A) + ID_B \times \delta(B) + ID_C \times \delta(C) +$$

$$+ ID_D \times \delta(D) + ID_E \times \delta(E) + ID_F \times \delta(F) =$$

$$= \frac{1}{12} + \frac{1}{12} + \frac{5}{6} + \frac{1}{6} + \frac{11}{4} + \frac{3}{4} = 4.67$$

Family Inherited Mutation Indicator. Given the individual inherited mutation indicators of all of the youngest members of a family, the family inherited mutation indicator consists on the addition of all of them, as follow:

$$FM_I = \sum_{\forall l | l \text{ is leaf}} M_I(l) \tag{8}$$

Family Integration Indicator. At this point we have provided ways of estimating different indicators which give us information about the mutation state and the affection rate in the family tree structure. As in the case of the global indicators, both kinds of indicators can be combined to set up a family integrator indicator relating the mutations and affectations evolution of the individuals over the tree, which provide a better indicator for representing the disease spreading in a given branch of the family.

Our proposal is based on a weighted average operator [5], thus given the individual affectation and mutation indicators of a family, the family integrator, FI is as follows:

$$FI(m) = f(FA_I, M_I) = \alpha FA_I + \beta M_I \tag{9}$$

Where:

- $\alpha, \beta \in [0, 1]$ are weights expressing the importance of affectation and mutation correspondingly, and $\alpha + \beta = 1$.

The selection of one value for α and β could be experimentally set according to the problem domain.

3.3. Extension to Multi-Rooted Pedigrees

Along this section, we have assumed that the pedigree has a unique root (see Figure 1), but there are situations where the specific family member who is being considered for the study has information from both, mother's and father's ancestors, like E in Figure 6 top. In this situation, a topological sorting solution is required. Topological sorting concerns the definition of the order in which nodes should be traversed in a complex graph.

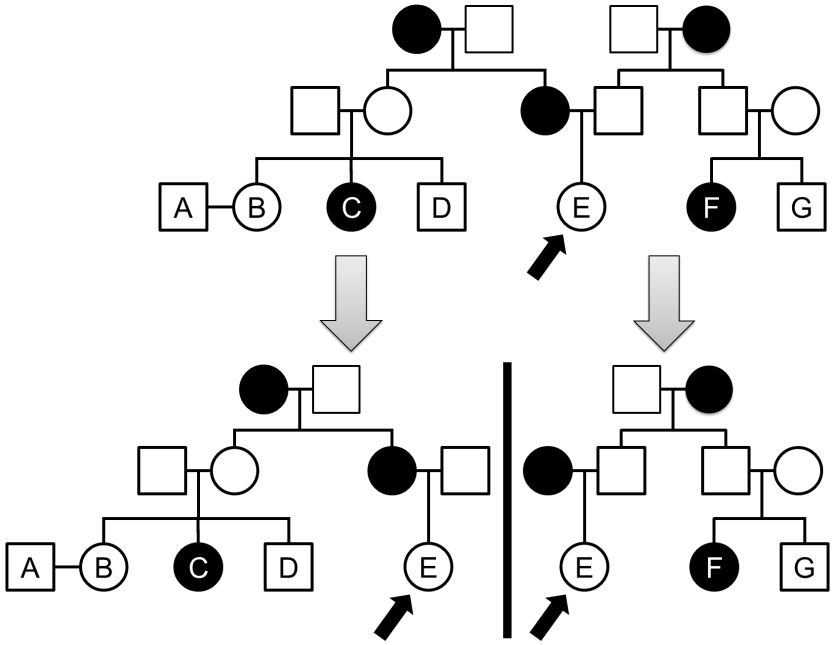

Figure 6: Tree example with multiple ancestors root.

In our case, instead of forcing a single list of ancestors, we propose the division of the multi-rooted pedigrees into single-root pedigrees like the ones we have been working until now, evaluate the indicators in each single-root tree, and provide the results in a confidence interval. The extremes of the

interval are the maximum and minim values obtained.

In order to do that, we first select all the family members without ancestors (pedigree's roots), and create from them new, single-rooted pedigrees. Obviously, we avoid the repetition of the same pedigree that could be created by selecting the husband and the wife of the same root. They are considered as a unique entry point to the pedigree. For example, if we look again at Figure 6, we can see that we have four family members from the top pedigree's roots. Since the four root candidates represent two marriages (two real roots), just one of them for each relation is necessary. Two pedigrees like the ones at the bottom of Figure 6 are created.

Then the indicators are estimated for each individual pedigree and instead of providing a final value, a value interval is provided. By this means, we assert that the family members that are the common branch of the different pedigrees have at least the worst estimated value and at maximum the best one.

4. Case Studies

This section shows real examples of pedigrees appropriately modified as case studies to visualize the utility of the indicators introduced in this paper. The first examples are simple and have been artificially generated to demonstrate the benefits of using structured data-based indicators. Figure 7 shows three times the same tree structure where changes among trees rely on the position where an affected node is detected. The node target of the study is

Case	A_g	M_g	P_g	A_I
Figure 7.a	0.2	0.4	0.5	5
Figure 7.b	0.2	0.4	0.5	2
Figure 7.c	0.2	0.4	0.5	1

Table 2: Global indicators and A_I according to where the affected node is situated within the tree structure

at the bottom, and it is highlighted inside a dashed box. As Table 2 shows, statistics-based indicators do not change independently on who the affected node is. Conversely, the individual inherited indicators depend directly on where the affected node is placed in the family.

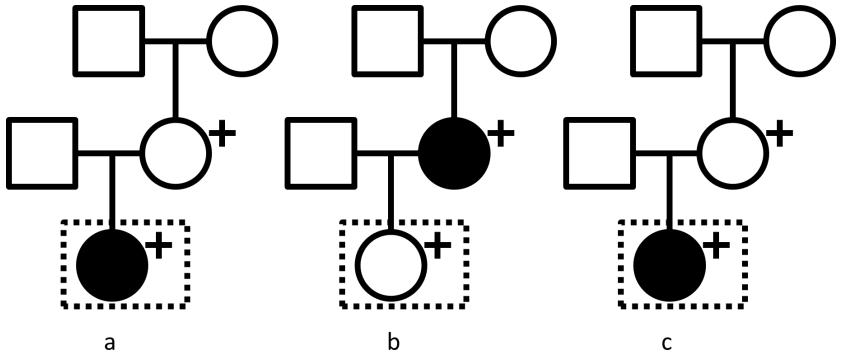

Figure 7: Figures a, b and c present the same family tree except for the location of their affected node.

Second case study is provided in Figure 8. It shows a simple family tree (Figure 8.a) to which more nodes are gradually added (Figures 8.b and 8.c). The node under study is highlighted inside a dashed box. The amount of mutated/affected nodes is not increased from one tree to the other.

As Table 3 shows, A_g and M_g indicators decrease their value quickly when more nodes are added. P_g is static because, as previously exposed,

the amount of mutated/affected is not modified. Regarding the structured data-based indicators, A_I shows a similar behavior as A_g and M_g, but in this case, it decreases proportionally slower. In contrast, M_I does not follow the same pattern as the other indicators. The reason rely on the exploration rules applied (Section 3.2.2). Figure 8.a follows Rules 9 and 7 so they only explore the father and the grandmother of the specified node. Then, in Figure 8.b, both father and mother have no information about mutations (neither + or -) so both are explored following Rule 11, and Rule 4 is used when the grandfather is explored. Finally, in Figure 8.c, since both parents contain the mutated gene, Rule 5 is triggered and next, when exploring the grandparents, we use again Rule 11.

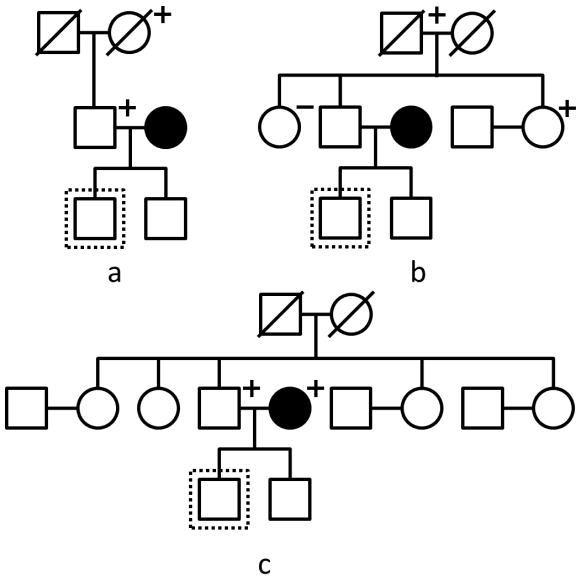

Figure 8: Figures a, b and c present a similar family tree structure except for the amount of brothers/sisters in the second era and the location of the mutated nodes. The node under study is highlighted inside a dashed box.

Case	A_g	M_g	P_g	A_I	M_I
Figure 8.a	0.16	0.33	0.5	1	1
Figure 8.b	0.11	0.22	0.5	0.33	0.25
Figure 8.c	0.08	0.15	0.5	0.2	0.5

Table 3: Indicators of the tree structures in Figure 8.

About the family inherited affectation indicator, we expect to behave similarly than the global indicators since they summarize the information of the whole family. Thus let us suppose the pedigrees depicted in Figure 9. The first pedigree, Figure 9.a, is a simple tree which has been modified so it does not include any of the genetic nor affectation information. Since it does not contain any trace of information, the results are zero for all of the indicators. In Figure 9.b, affectation information is included, so the global affectation indicator grows until a 0.39. Figure 9.c includes genetic information; therefore, the global mutation indicator is 0.22. If both are considered (Figure 9.d), the global penetration escalates until 1.75, which represents a high level of penetration given that there are more affected nodes than mutated. The case study of Figure 9.e consists in a highly complex multi-rooted pedigree. All of the global indicators obtained are summarized in Table 4, together with the family inherited indicators. It's easy to see the evolution of the values according to the cases and the mutated/affected elements introduced, as in the global indicators. The family integrator indicator F_I in this case has been estimated using '$\alpha = \beta = 0.5$ and, as can be seen, the value depends directly in the era where the affected node is detected.

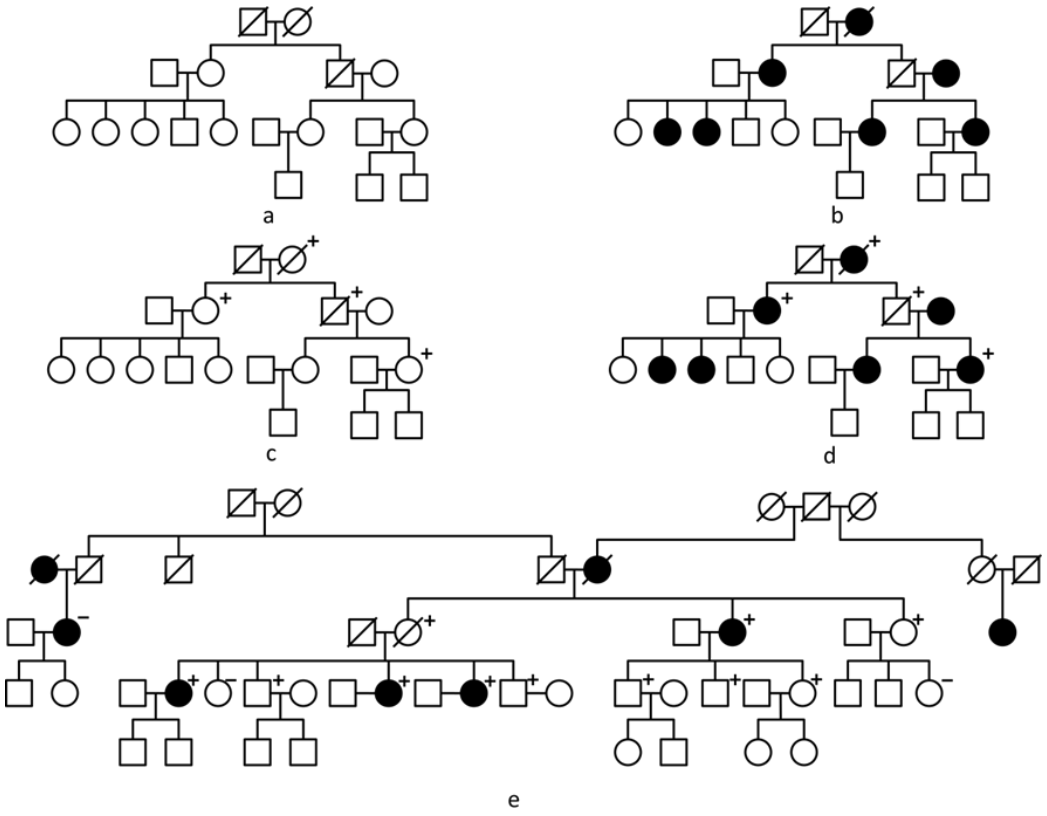

Figure 9: Case studies. (a) Pedigree free of disease. (b) Pedigree with some of the members affected. (c) Pedigree with with members who are carriers of the genetic mutation. (d) Combination of (b) and (c). (e) Case study of a real multi-rooted pedigree.

5. Experimental Results

The implementation of our methodology has been done in Java and provided as a plugin for eXiT*CBR [6], a tool for case-based medical diagnosis support. The GUI is simple (see Figure 10), so that the user provides the information about:

- The indicators to be computed.

Case	A_g	M_g	P_g	FA_I	FM_I	FI
Figure 9.a	0	0	0	0	0	0
Figure 9.b	0.39	0	0	5.8	0	2.90
Figure 9.c	0	0.22	0	0	5.17	2.58
Figure 9.d	0.39	0.22	1.75	5.8	5.17	5.48
Figure 9.e	0.16	0.22	0.78	[4.33,3.56]	[4.5,5.5]	[4.47,4.03]

Table 4: Comparative between the different statistics-based indicators against the structured data-based (family) indicators regarding the case studies.

- The comma separated value file with the list of the GED files corresponding to the families to be analyzed.
- The directory where the GED files are located.
- The name for the results file.

When choosing the inherited mutation option, both family indicators introduced in this paper are calculated. The penetrance indicators are computed as a post process, if both, affectation and mutation are chosen, since it is derived from them.

To test the benefits of our methodology we have used our breast cancer data which consists of 347 families (GED files)[1]. Among all of the individuals, we have clinical information from 553 members to validate the results obtained; that is, we know if the individual has suffered or not the illness (155 and 399 correspondingly).

[1]Unfortunately the data is not public due to medical constraints. Any researcher interested can send an e-mail to the authors, to ask for permission to the medical staff.

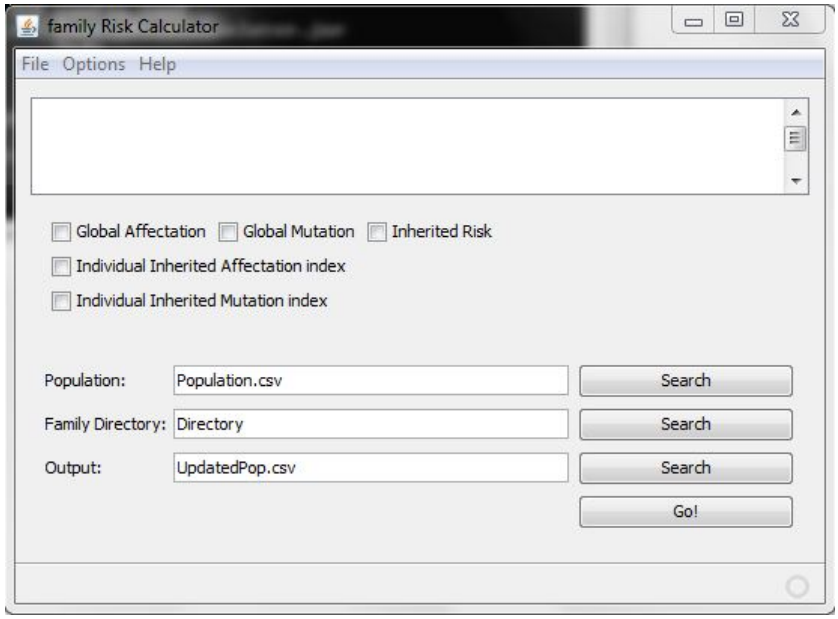

Figure 10: GUI for the implemented methodology.

All of the indicators have been computed for the 553 individuals. Regarding the parameters of the family integration indicator they are set to 0.5 each (i.e., $\alpha = \beta = 0.5$ in the equation from Equation 9). To measure the experiments we use the confusion matrix and the true positive and true negative rates that can be derived from it. The confusion matrix relates the number of false positive (FP, healthy people detected as ill sufferers), false negatives (FN, ill sufferers predicted as healthy), true positives (TP, ill sufferers predicted correctly) and true negatives (TN, healthy people with right prognosis). The true positive (tp) rate relates the number of ill sufferers detected with the indicators regarding the total of known patients (i.e. TP / (TP+FN)); the true negative (tn) rates the number of individuals

with indicators given zero information with all healthy people (that is, TN / (TN+FP)). The threshold value to predict that a person could be ill sufferer is 0; that is, a person with an indicator value higher than zero is considered as a candidate to develop the illness. This threshold should be revised in a future work with the feedback of physicians.

Table 5 provides the results obtained. Analyzing the results, we observe that the global affectation indicator is not able to discriminate between patient suffering the illness and healthy people, labeling all members as possible candidates for suffering the illness. The individual inherited affectation indicator is able to discriminate up to 83.23% of the ill people; although the tp-rate does not achieve the 100% and this is a costly situation. An interesting result is the family inherited affectation indicator that results in an equivalent behavior than the global affectation indicator, as could be expected, since it covers all members of the family.

Regarding mutation, it is a weak indicator for illness risk assessment, as some physicians anticipate. There are a lot of people that could have a given mutation but it is not a definitive factor for developing the illness.

However, it is important to highlight the fact that the individual indicator (M_I) reach up to the 91.98% for tn-rates, outperforming the global indicator in almost the double. Analyzing in detail mutations, we know that there were 36 proven mutations in the family tree; the global mutation indicator identifies 249 possible individuals who can have the mutation, close to the family inherited mutation indicator (240), while the individual inheritance

Indicator	FP	TP	FN	TN	tp-rate (%)	tn-rate (%)
A_g	399	155	0	0	100.00	0.00
M_g	187	62	93	212	40.00	53.13
P_g	187	62	93	212	40.00	53.13
A_I	112	129	26	287	83.23	71.93
M_I	32	28	127	367	18.06	91.98
FA_I	399	155	0	0	100.00	0.00
FM_I	182	58	97	216	37.42	54.27
P_b	399	155	0	0	100.00	0.00

Table 5: Experimental results.

only 60, thanks to the inheritance rules designed. So the rules seem to play a key issue on appropriately assigning mutation risks to individuals.

Concerning the penetration indicator, it is dominated by the mutation results, while the branch penetration by the affectation indicator. These results require further research.

Finally, it is important to highlight that those indicators represent a first step towards the automation of the evaluation of the information stored in pedigrees, so that they can be available to be combined with other clinical data in order to achieve more accurate prognosis.

6. Related Work

There are some methods that deal with inheritance information due to the interest of evaluating the disease risk factors by the aggregation of cases (i.e. persons suffering the illness). However, most of the works concerns the use of pedigree information in order to determine features of a particular population under study. In [7], for example, several approaches are studied

to determine the right number of contributions that should be considered from the ancestors in a given population. Most of the analyzed methods are based on the inverse proportion to the addition of the square of all of the features under study. Our approach also takes into account this inverse proportion, but in this case, instead of using the square, we exponentially modify the information according to the distance of the ancestor to the current population.

Another interesting work is [8] that apply aggregation methods to evaluate cardiological risks taking into account fathers/soon exponential relationships. Our proposal also includes an exponential relation, but taking into account the complexity of the whole family, which includes hierarchical relationship of different generations, and different family trees as a consequence of marriages. Thus, our work contemplates different sources of risks, depending on the different paths that can be followed in a graph-representation of the family.

The most interesting insight of the [8] work is how the combination of the different induced information from the same pedigree can be performed. In particular, the authors propose several functions to combine risk and incidence factors. Thus, we leave for a future work to use this kind of aggregation methods, as well as other methods coming from the veterinarian field [9] and decision theory [5].

Regarding breast cancer diagnosis BRCAPro [10] also deals with the problem of analyzing ancestors in order to find a possible mutation. BRCAPro

is limited in depth, analogously to the case of [8], but in this case up to 2 degrees of relationships are allowed (grandfather/mother, father/mother and son). BOADICEA [11] and Tyrer-Cuzick [12] are similar-purpose tools.

An interesting work is [13] which describe the complexity of deriving the individual risk from the population risk. Particularly the authors comment three factors to take into account: having a family history or not, contribution of the population model in the family predisposal, and how environmental risk factors affect individuals. Based on these factors, the authors propose a mathematical model to obtain individual values. Our work is related to provide support to such kind of mathematical models; in this first work, we are proposing the measure of the information contained in the family, but as a future work, more complex models as the ones proposed in [13] or [14] that merge population (statistics) models should be contemplated.

Finally, it is important to note that current pedigree tools, as PyPedal [15] are including different methodologies to analyze the data included, so as to summarize, as for example in the case of PyPedal, in eight different measures (similar as the ones previously related). Then, we should expect in a future that thus tools also incorporate the facility of analyzing pedigrees according to user-provided methods, or domain-dependent methods. Our method could be one of them.

7. Conclusions

The need of dealing with pedigree analysis is a must when dealing with inherited illnesses. However, most of the approaches to interpret data on family histories are statistics based on the number of individuals in the family instead of their relationships. In this work we presents a new way of computing family risk based on indicators that takes into account the structure of the family, as physicians use to do.

We provide a methodology based on three statistics-based indicators (global affectation, global mutation, and global penetration indicators), plus five structured data-based indicators (individual inheritance mutation, individual inheritance affectation, family inheritance affectation and mutation, and family integration indicators). To establish relationships among family members, a rule-based algorithm is provided for simple pedigrees (tree-like structures), while an interval based solution is provided for more complex situations (graphs due to second marriages, etc.).

We have illustrated our methodology on several case studies. Moreover, our method has been implemented and integrated in a case-based reasoning tool that supports medical decision making. The experimentation has been carried out in a breast cancer diagnosis domain, showing that the inheritance information computed with our indicators can be more discriminative than statistics approaches.

Having a way to compute indicators, the next step of our research is to discuss the results with physicians to enrich our methodology. This future

work comprehends the combination of statistics and structured data based indicators (we have only tackled combinations among structured data based indicators). Our ultimate goal is to provide to the physicians with a powerful tool that supports their decision making taking into account the interrelationships among population values and individual information, that it is still an open problem.

References

[1] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, J. J. Mulvihill, Journal of the National Cancer Institute 81 (1989) 1879–1886.

[2] CyrillicSoftware, Cyrillic, http://www.cyrillicsoftware.com/, 2012.

[3] Family History Department , The Church of Jesus Christ of Latter-day Saints, The GEDCOM standard release 5.5, http://www.gedcom.net/0g/gedcom55/, 2012.

[4] G. Malécot, Les Mathématiques de l'Héredité, Mason & Cia Editeurs, 1948.

[5] V. Torra, Y. Narukawa, Modeling Decisions: Information Fusion and Aggregation Operators, Springer, 2007.

[6] B. López, C. Pous, P. Gay, A. Pla, J. Sanz, J. Brunet, Artificial Intelligence in Medicine 51 (2011) 81 – 91.

[7] D. Boichard, L. Maignel, E. Verrier, Genet Sel Evolution (1997) 5–23.

[8] O. O. Aalen, Biometrics 47 (1991) 933–945.

[9] C. L. Battaglia, Journal of Veterinary Behavior: Clinical Applications and Research 3 (2008) 183.

[10] D. Berry, G. Parmigiani, J. Sanchez, J. Schildkraut, E. Winer, J. Natl. Cancer Inst. 89 (1997) 227–238.

[11] A. C. Antoniou, P. P. D. Pharoah, P. Smith, D. F. Easton, British Journal of Cancer 91 (2004) 1580–1590.

[12] P. J. Cuzick, IBIS breast cancer risk evaluation tool, http://www.ems-trials.org/riskevaluator/, 2012.

[13] R. A. Kerber, Genetic Epidemiology 12 (1995) 291 – 301.

[14] C. Cannings, M. H. Skolnick, K. de Nevers, R. Sridharan, Computers and Biomedical Research 9 (1976) 393–407.

[15] J. B. Cole, Computers and Electronics in Agriculture (2007) 107–113.