Journal of Biomedical Informatics 46 (2013) 710-720

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

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

ARTICLE INFO

Article history: Received 7 January 2013 Accepted 7 June 2013 Available online 15 June 2013

Keywords: Decision support Structured data Feature construction Modeling Medical applications Knowledge-based system

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, a knowledge-based system (KBS) 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

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 systems 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 been experimentally tested with breast cancer diagnosis data. The results show the feasibility of our methodology. © 2013 Elsevier Inc. All rights reserved.

> 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 KBSs 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 5 case studies are provided, and in Section 6 the experimentation performed so far is shown and discussed. Then, in Section 7 we expose some related work and, finally, we end the paper in Section 8 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 these sections, records repre-







^{*} Corresponding author.

E-mail addresses: pablo.gay@udg.edu (P. Gay), beatriz.lopez@udg.edu (B. López), albert.pla@udg.edu (A. Plà), jordi.saperas@udg.edu (J. Saperas), carles.pous@udg. edu (C. Pous).

^{1532-0464/\$ -} see front matter © 2013 Elsevier Inc. All rights reserved. http://dx.doi.org/10.1016/j.jbi.2013.06.003

sent 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 Fig. 1, the representation of a GEDCOM 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 does not have 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 KBS 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. 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.

Fig. 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 KBS to support medical decision making, as presented at the end of this 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.



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

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 Fig. 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:

$$M_{\rm g} = \frac{M}{T} \tag{2}$$

where

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

Following the example of Fig. 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 Fig. 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 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 Download English Version:

https://daneshyari.com/en/article/10355490

Download Persian Version:

https://daneshyari.com/article/10355490

Daneshyari.com