
Identification of Relevant Knowledge for Characterizing the Melanoma Domain

Ruben Nicolas¹, Elisabet Golobardes¹, Albert Fornells¹, Susana Puig²,
Cristina Carrera², and Josep Malvehy²

¹ Grup de Recerca en Sistemes Intel·ligents, Enginyeria i Arquitectura La Salle, Universitat Ramon Llull, Quatre Camins 2, 08022, Barcelona, Spain

² Melanoma Unit, Dermatology Department, Hospital Clinic i Provincial de Barcelona, IDIBAPS, Barcelona, Spain
{rnicolas, elisabet, afornell}@salle.url.edu,
{spuig, ccarrera, jmalvehy}@clinic.ub.es

Abstract. Melanoma is one of the most important cancers to study in our current social context. This kind of cancer has increased its frequency in the last few years and its mortality is around twenty percent if it is not early treated. In order to improve the early diagnosis, the problem characterization using Machine Learning (ML) is crucial to identify melanoma patterns. Therefore we need to organize the data so that we can apply ML on it. This paper presents a detailed characterization based on the most relevant knowledge in melanomas problem and how to relate them to apply Data Mining techniques to aid medical diagnosis in melanoma and improve the research in this field.

Keywords: Health Information Systems, Knowledge Management and Decision Support Systems, Melanoma domain, Computer Aided Systems.

1 Introduction

Melanoma is now one of the most social interesting cancers because it is more frequent in our society and affects people of any age. According to the *American Cancer Society* although is not the most common skin cancer, it is which causes most deaths. This increase, caused by solar habits, makes crucial the early diagnosis, even more if we analyze that this cancer is mortal in approximately twenty percent of cases [1,2] and prompt diagnosis permits practically a secure regain.

Nowadays domain experts works with some plain data bases with disjoined information that does not permit roomy experiments to identify melanoma patterns. In fact, these days diagnosis is not aided for computer systems and requires checking several reports, from different specialists, to give a unified prognostic. Dermatology experts from *Hospital Clinic i Provincial de Barcelona* (HCPB) want to do statistical analysis of its data, and characterize new melanoma patterns.

One way to improve the early diagnosis is to use Knowledge Management and Decision Support Systems based on the statistical results of the cancer information. This imply a Data Mining (DM) problem were we have to analyze high dimensional data, with uncertainty and missing values, to extract the patterns for aid decision making [3]. Some of these techniques have been proved in other kinds of cancer, like the

breast one [4], with results that permits promising investigation. But this technology needs a good data characterization and organization to work and this is non-existent in its domain. To create it we have analyzed data from more than three thousand melanoma patients with information from reports of dermatologists, oncologists, surgeons, pathologists, and other specialists that work at HCPB. Some of these data are used in international studies that analyze specific aspects of the domain. But medical researchers want to study it on the whole to assess if clear patterns are reachable. This aim makes these data really unique and valuable for its variety and completeness.

Then, we have to propose a data organization in melanoma domain to find out which kind of information is relevant and how it is related. This proposal will be the base line for building a DM tool to help experts in melanoma research. Now we want to apply clustering techniques to divide the domain, as is intended previously with less data [5], and use Case-Based Reasoning to support the diagnosis.

The paper is organized as follows. Section 2 analyzes the framework in the field we would like to contribute. Section 3 describes the melanoma domain characterization. Finally, in sections 4 and 5 we discuss about the problem and mark the future work.

2 Related Work

Current works in the melanoma study treat the issue as a sectioned one. We could found works which use only some clinical and pathological data, but with a large number of cases [6], or that studies particular types of information over specific individuals [7,8]. This kind of reference is based on partial databases which are centred on concrete information that does not use the entire data of the domain. A support system that permits us to put it at stake could open new targets in the future using DM. Since early diagnosis of melanoma is the most important factor in the progress of the disease, the diagnostic accuracy is of major interest; is required to establish concise diagnostic criteria based on the clinical information. If we want to apply computer aided systems for getting these objectives, we need a complete and consistent relational model.

3 Melanoma Framework

To apply DM techniques in order to extract patterns from the complete data of the melanoma domain we have to unify the data from different plain databases, from diverse medical specialities, medical information from non electronic support and new data from specific diagnostic machines. This section describes the complexity and constrains of melanoma domain. But, a basic conceptual explanation of the domain is not enough to represent all the relations, specific restrictions, and rules to follow in data insertion, maintenance, and search. Then, with all the collected information we have planed a relational model shown in Fig.1. This model permits all the storage capacities we need, respect the relations and prevent the insertion errors to avoid redundancies. Experts consider that the next aspects explain the domain: 1) person and family, 2) generic medical information, 3) tumors, 4) metastasis, and 5) controls and studies.

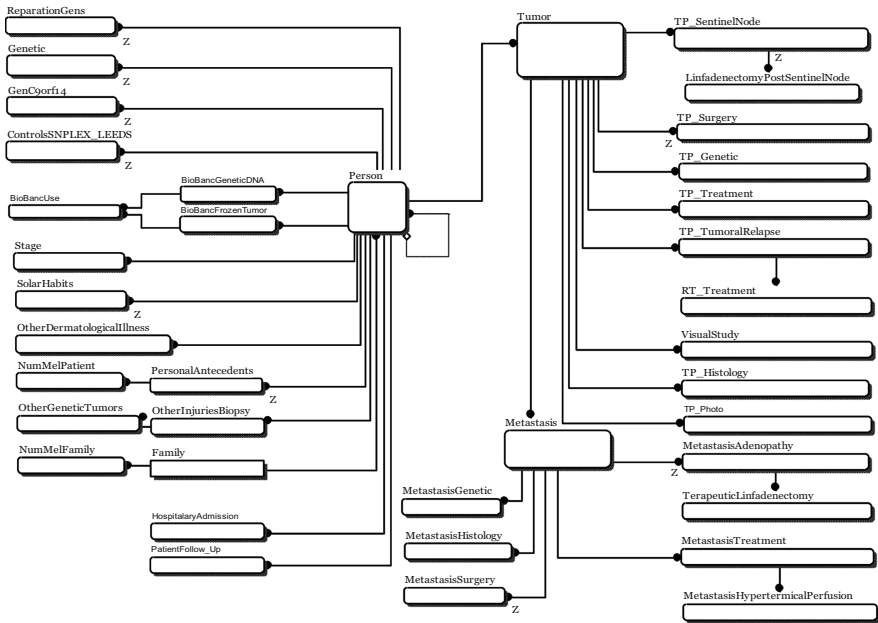


Fig. 1. Relational Model of Melanomas Domain

Person and Family. This aspect summarizes the personal data and antecedents of a person. We note that there are different kinds of person: melanoma patients, familiars of these patients and control cases (healthy people). Patients are also related with their DNA samples and the hospitalary admissions and follow-up that permits to check the illness evolution. Emphasized information about the patient is to know the stage of the illness in different moments (indexed per date), the solar exposure, other dermatological illness, and its biopsies. Another important data to take into account is the familiar relations because it is possible a familiar development of the melanoma.

Tumors. This entity is related with a unique patient, but one patient could have more than one tumor. In addition, this entity has relation with its histologies, treatments applied, genetical information, surgery data, sentinel node, and metastasis. We also have the relation between the tumor and its different images.

Metastasis. This is a concept that depends on a tumor. Each tumor could have different metastasis. The general information is at the same time in relation with histologies, treatments, surgery, genetics, and lymph node metastasis that is included with the addition of linfadenectomy as a particular idea of a metastasis.

Controls and Studies. These aspects refer to the research studies developed by experts in the identification of patterns. A study is the analysis of a certain aspect of a patient. This concept is strongly linked with the person relation because they are who take part in the studies.

4 Discussion

The aim of our current research is to obtain clear melanoma patterns based on all the data existent in this domain. We would like to clusterize the complete information in order to create patterns of diverse characteristics that permit an easy prognostic from different kinds of melanoma even for non-expert doctors. We have also the target of obtaining a good ontological description that allows a better use of CBR techniques to aid medicians in diagnostic accuracy. But all this aims need a previous step that is the characterization of the domain that permits the desired investigation.

The characterisation exposed in section 3 structures melanoma domain and permits not only the target of allowing subsequent work of aid melanomas early diagnosis and accuracy, but it has a suitable insertion of information. This is important because we have found during the process that the previous data have inconsistencies or not related information that come from a bad pick up of it. The incorporation of triggers permits to control this situation and, in addition, filter the existent one in the migration process. Then, at the end of the process, we have obtained a new database completely consistent and with well related information.

We would like to note that the elaboration of the model adds new data obtained from paper reports and diverse databases not unified since now. This situation permits to prepare complete sets of raw data that allows the obtaining of melanoma patterns and statistics for medical research.

5 Conclusions and Further Work

Melanoma is a dermatological cancer with an important impact in our society. Medical researchers in this field want to use techniques to aid its diagnosis. The first step to permit the use of these techniques is the characterization of the domain in order to use the complete data in its studies. The definition proposed permits the creation of an application that allows: 1) An easy data introduction from the specialists, even during attendance work. 2) Knowing the trustworthiness grade of data in order to consider or not low reliable data in certain studies. 3) Rapid creation of new experiments and statistical results in medical research. 4) Study relations between different melanomas information in the bosom of a family. 5) Apply Data Mining Techniques in order to aid specialist to obtain optimal descriptions of patterns to improve the early diagnosis.

Nowadays, our work is focus in the final implementation of the application and the migration of the previous data. After this migration, we could apply Data Mining techniques to create a complete Knowledge Management and Decision Support System *ad hoc* to melanoma's domain that permits to help the medical researchers in the prognostic of this illness. This work are planed in two phases, the first one the extraction of patterns and explanations from this complete data, enlarging previous works that use parts of it [5]; and the second that is the implementation of a hierarchic reasoner that permits to aid the diagnosis by analysing the different kinds of information in separated stages to conclude an unified result.

Acknowledgements

We thank the Spanish Government for the support in MID-CBR project under grant TIN2006-15140-C03 and the Generalitat de Catalunya for the support under grants 2005SGR-302 and 2008FI_B 00499. We thank Enginyeria i Arquitectura La Salle of Ramon Llull University for the support to our research group. The work performed by S. Puig and J. Malveyh is partially supported by: Fondo de Investigaciones Sanitarias (FIS), grant 0019/03 and 06/0265; Network of Excellence, 018702 GenoMel from the CE.

References

1. Gutiérrez, R.M., Cortés, N.: Confronting melanoma in the 21st century. *Med. Cutan. Iber. Lat. Am.* 35(1), 3–13 (2007)
2. Tucker, M.A., Goldstein, A.M.: Melanoma etiology: where are we? *Oncogene* 22, 3042–3052 (2003)
3. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*. Elsevier, Amsterdam (2006)
4. Fornells, A., Golobardes, E., Bernadó, E., Martí, J.: Decision Support System for Breast Cancer Diagnosis by a Meta-Learning Approach based on Grammar Evolution. In: 9th ICEIS, pp. 222–227. INSTICC Press (2006)
5. Fornells, A., Armengol, E., Golobardes, E., Puig, S., Malveyh, J.: Experiences Using Clustering and Generalizations for Knowledge Discovery in Melanomas Domain. In: ICDM 2008. 7th Industrial Conference on Data Mining. LNCS, pp. 57–71. Springer, Heidelberg (2008)
6. Balch, C.M., et al.: Prognostic Factors Analysis of 17,600 Melanoma Patients: Validation of the American Joint Committee on Cancer Melanoma Staging System. *J. Clin. Oncol.– Am. Soc. Clin. Oncol.* 19, 3622–3634 (2001)
7. Puig, S.: Role of the CDKN2A locus in patients with multiple primary melanomas. *J. Clin. Oncol.* 23(13), 3043–3051 (2005)
8. Malveyh, J., et al.: On behalf of the International Dermoscopy Society Board members. Dermoscopy report: Proposal for standardization Results of a consensus meeting of the International Dermoscopy Society. *J. Am. Acad. Dermatol.* 57(1), 84–95 (2007)