CARA-D: Data Elements for a Computer based Cancer Risk Assessment System

Gi-Chul Yang and Haeng-Un Oh

School of Information Engineering
Mokpo National University
Mokpo, KOREA
gcyang@mokpo.ac.kr
huoh@korea.com

Abstract: Data elements are important part of a computer based cancer risk assessment system. The selection of the data elements are more important for the system built based on Case Based Reasoning (CBR) technology. The system CARA is a computer based cancer risk assessment system that adapts CBR technology. The data elements and an overall architecture that can assure high performance of the CARA are described in this article.

1 Introduction

The CARA (CAncer Risk Assessment) system is a computer based cancer risk assessment system, which can help people by finding their chances of having an inherited susceptibility to cancer. Doctors and researchers, also, can use the CARA to find unknown relationships and factors in between certain cancer and genetic information related to a family history.

Even with the current medical technology, early detection is the best way to cure the cancer. In order to detect cancer as early as possible, regular periodical tests are important. However, people do not get the test regularly due to the lack of medical resources and their own personal reasons. So, it would be very helpful if we have a system, which can tell us the possibility of having cancer just by providing our cancer related medical information. The CARA is such a system.

This research was financially supported by the Ministry of Commerce Industry and Energy (MOCIE) and Korea Industrial Technology Foundation (KOTEF) through the Human Resource Training Project for Regional Innovation.

There are well-studied criteria for cancer risk assessment [1,2] and the CARA can be used to prove the correctness of those studies by applying the real data directly. The CARA adopts the case-based reasoning (CBR) technologies and it can provide accuracy of the result without distortion of the data analysis that often occurred during the mathematical formula calculation. During in each step of the mathematical formula calculation performed in other cancer risk assessment methods [3,4,5], small errors can be accumulated and those accumulated errors make big differences at last. In order to produce an accurate result through the CARA, the input data provided to the CARA should be accurate and the strong relationships to the cancers are required. The efficiency and the accuracy of the CARA depends heavily on the data elements that the CARA using. The necessary and important data elements and their data structures along with the overall architecture for the CARA are described in this paper.

Overall architecture of the CARA is explained in section 2 and the input data elements are described in section 3. In section 4, the CARA interface is shown and the paper is concluded in section 5.

2 The CARA System

The CARA is designed based on the CBR technology and the Internet. It is a good idea to build a system like the CARA on the Internet since the CARA requires large number of cases and the Internet can provides an easy way of collecting such data. Also, it is easy to access the system on the Internet for the users. By adapting the CBR technology, the CARA will be able to reduce the accumulated errors that often cause the problems in other methods[3,4,5]. The general architecture of the CARA depicted in Figure 1 and more details of the CARA architecture are described in [6].

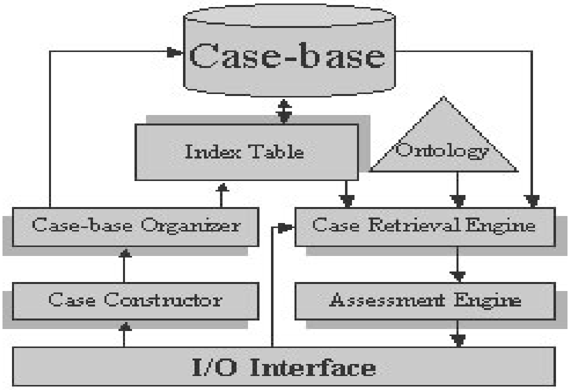


Figure 1. The CARA System

The case constructor accepts the information from the user through the interface and constructs a pedigree. The Case-base Organizer builds and reorganizes the Index Table and stores the input case into the Case-base. The Case-base holds all the cases submitted to the system. The CARA is a system based on case-based reasoning technology. Therefore, case retrieval engine is an essential component of the CARA. The CARA used Index Table, which is explained in [7] to retrieve the cases efficiently. The details of the case retrieval engine are depicted in Figure 2.

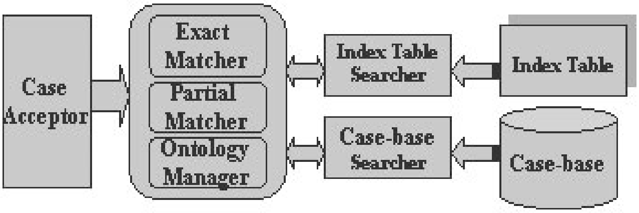


Figure 2. Case Retrieval Engine

The case retrieval engine has abilities of exact matching as well as partial matching. Hence, the case retrieval engine can retrieve exactly same pedigree as well as partially similar pedigrees according to the user's request. This ability makes the CARA a realistic system.

In case of the partial matching, there are structural partial matching and semantic partial matching. The structural partial matching just considers the structure of the pedigrees but the pedigrees may be different if the nodes contain different information. Hence, the ability of semantic partial matching is required to cover those cases. Index table is needed for speed up the case retrieval process. The case-base organizer builds index table and store the incoming case to the case-base.

The assessment engine performs cancer risk assessment based on the result of case retrieval engine. And the overall result of the cancer counseling is given to the user through the i/o interface. Ontology is needed for semantic partial matching. For example, a node in a family tree represented as a cancer-affected node and a node in the other pedigree is represented as a breast cancer affected node. Then, even though the two pedigrees are having same structure those are not matched by exact matching, since those pedigrees are structurally same but semantically different.

3 Data Elements for the CARA

In order to estimate an individual's hereditary cancer risk, it is essential to summarize family history in the form of a pedigree [2]. The CARA is designed based on the Case Based Reasoning (CBR) technology and the pedigree is the backbone structure of the case for the CARA. There are pedigree examples in Figure 3.

The pedigrees in Figure 3 are simulated pedigrees and the arrow ↗indicates the proband and ● indicates the affected female. The pedigrees in Figure 3 just show the structure of the family members and the affected individuals. The CARA needs more information on each node of the pedigree.

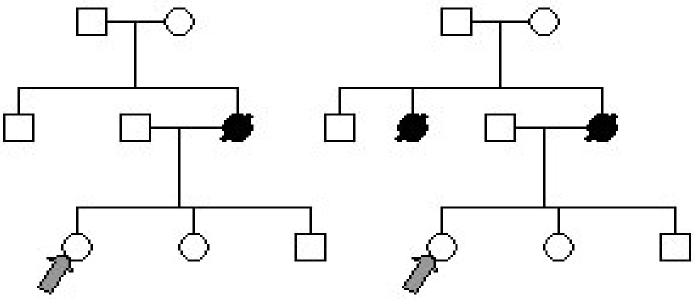


Figure 3. Example Pedigrees

The data elements for the CARA are decided according to the information provided from National Cancer Center (NCC) of Korea and cancer risk assessment related research results such as [1,2,8,9,10]. There are more than 26 questions as well as dietary habits and some questions specifically for woman are provided in the questionnaire of NCC. Following topics are included in the questionnaire of NCC.

- Personal medical history besides cancer
- Personal history of cancer
- Personal history of medication
- Relatives history of cancer
- Relatives medical history
- Hepatitis type B inoculation
- Smoking
- Drinking
- Exercise etc.

Following are factors suggesting inherited cancer risk in [1,2].

- Clustering of the same type of cancer in close relatives.
- Unusually early age of cancer onset.
- Two or more primary cancers in a single individual.
- Evidence of autosomal dominant inheritance.
- Bilateral cancer in paired organs.
- Multifocal cancer.
- Patterns of cancer in the family that are associated with a known cancer syndrome

Also, followings are information typically included in a cancer family history explained in [2]:

- Both maternal and paternal relatives.
- Race, ancestry, and ethnicity information for all grandparents.
- Information about seemingly unrelated conditions, such as birth defects or other nonmalignant conditions of children and adults that may aid in the diagnosis of a cancer susceptibility syndrome.
- A minimum of 3 generations.
- Notation of adoption, nonpaternity, consanguinity, and use of assisted reproductive technology (e.g., donor egg or sperm), when appropriate.

Typically collect the following information for any relative with cancer [10]:

- Type of each primary cancer.
- Age at diagnosis for each primary cancer.
- Where the relative was diagnosed and/or treated.
- If the individual is still living, current age; if deceased, age at death and cause of death.
- Carcinogenic exposures (e.g., tobacco use, radiation exposure).
- Other significant health problems.

And for any relative not affected with cancer, collect the following information [10]:

- Current age or age at death.
- If deceased, cause of death.
- History of any surgeries that reduce the risk for cancer.
- Whether routinely screened for cancer.
- Any nonmalignant features of the syndrome in question.
- Carcinogenic exposures.
- Other significant health problems.

According to the above information, we decided to use the following information as the main data elements for CARA.

- Personal medical history besides cancer
- Personal history of cancer
- Personal history of medication
- Personal history of acupuncturation
- Personal history of blood transfusion
- Personal history of residential environment
- Personal conditions of illness (symptoms)
- Personal dietary [food] habit
- Relatives history of cancer
- Relatives medical history
- Hepatitis type B inoculation (Personal)
- Smoking habit
- Drinking habit
- Exercise etc.

Each data element is composed with detailed subcategories such as any medical test or surgery (list is provided) for personal medical history besides cancer and so on. Also, information something about menstruation, birth experience, use of contraceptive pill, which related only for woman are included.

All the pedigrees consist with a minimum of 3 generations. And each node of the pedigree will have the following information as attributes. Age at diagnosis for each primary cancer, if the individual is still living, current age; if deceased, age at death and cause of death and so on. Notation of adoption, nonpaternity, consanguinity, and use of assisted reproductive technology are not included in the data element for the CARA since those are not the factor of genetic risk inheritance. Also, race, ancestry, and ethnicity information are not included in the data element for the CARA since those are rare and not easy to collect that information in Korea.

4. The CARA Interface

Currently, the CARA is being implemented on a PC based stand-alone system. However, the final version of the CARA will be able to run on the Internet. Fig. 4 is the current version of the CARA Interface for case construction. The main structure of the case for CARA is pedigree and the user can input their pedigree information through the table form of interface as in Fig. 4.

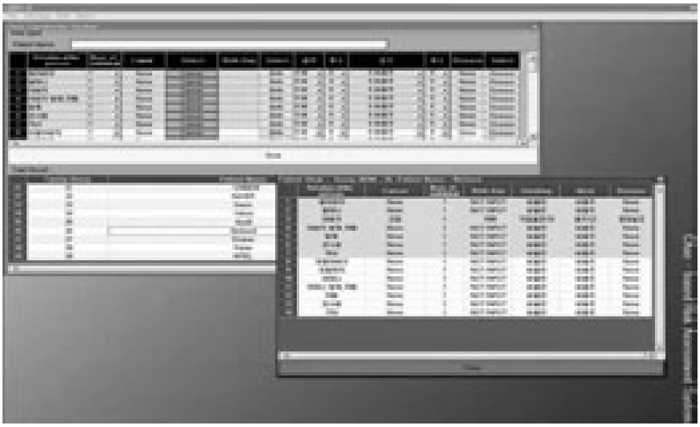
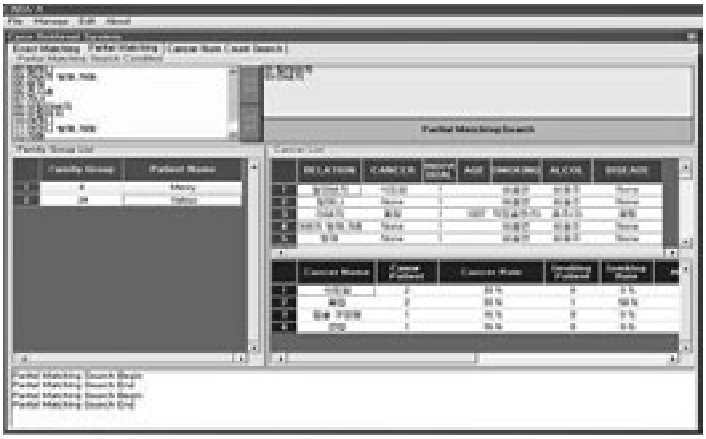


Figure 4. The CARA Interface for Case Construction.

The data elements decided in section 3 will be recorded through this table. The cases are stored in a database and can be retrieved through the interface shown in Fig. 5.



```
Partial Matching Search Begin
Partial Matching Search End
Partial Matching Search Begin
Partial Matching Search End
```

Figure 5. The CARA Interface for Case Retrieval

The CARA has the ability of exact matching as well as partial matching. In case of the partial matching, the percentages of the matching result are calculated and show as the statistics of the retrieval result. Hence, we can figure out the relation between certain cancer and the certain factors (data elements). This information is an important criterion to (genetic) assessment the risk of certain cancer.

5. Conclusion

The data elements for a computer based cancer risk assessment system are described in this article. The implementing computer based cancer risk assessment system is called CARA. The CARA is adapting CBR technology and the data elements selected and described in this article will be used by the CARA. The efficiency and accuracy of the CARA depends heavily on the data elements decided in this article. The data elements provided in this article are selected after thorough study of the previous research results and we believe that the decided data elements can provide the solid foundations of the CARA and can lead the accurate result of the CARA.

Rreferences

[1] Hampel, H., K Sweet, JA Westman, K Offit and C Eng, 2006. Referral for cancer genetics consultation: a review and compilation of risk assessment criteria. Journal of Medical Genetics.

[2] National Cancer Institute, Elements of cancer genetics risk assessment and counseling, ww.nci.nih.gov/cancertopics/pdq/genetics/risk-assessment, May 2006.

[3] Pharoah, P., et.al., 1997. Family history and the risk of breast cancer: a systemic review and meta-analysis. International Journal of Cancer.

[4] Benichou, J., et. al., 1996. Graphs to estimate an individualized risk of brest cancer. Journal of Clinical Oncology.

[5] Claus, E., et. al., 1990. Age at onset as an indicator of familial risk of breast cancer. American Journal of Epidemiology.

[6] Yang, G-C. and Park, J., 2007. CARA-A: An Architecture of Cancer Risk Assessment System based on Case-based Reasoning. APIS06, Kuala Lumpur, Malaysia.

[7] Yang, G-C. and Park, H., 2006. Representation of Family Tree for CARA. IMECS2006, Hong Kong.

[8] Emery, Jon, et. al., Computer support for interpreting family histories of breast and ovarian cancer in primary care: comparative study with simulated cases, Information in practice, BMJ Vo.321, July 2000.

[9] Evans, D. and Lalloo, F., Risk assessment and management of high risk familial breast cancer, Journal of Medical Genetics, Vo.39:865-871, 2002.

[10] Schneider K: Collection and interpretation of cancer histories. In: Schneider KA: Counseling About Cancer: Strategies for Genetic Counseling. 2nd ed. New York, NY: Wiley-Liss, 2001, pp 129-166.