Unification of heterogeneous data towards the prediction of oral cancer reoccurrence

Konstantinos P. Exarchos[1,2], Yorgos Goletsis[3], Dimitrios I. Fotiadis[1,*]

[1] Unit of Medical Technology and Intelligent Information Systems, Dept. of Computer Science, University of Ioannina, Ioannina, GREECE
[2] Dept. of Medical Physics, Medical School, University of Ioannina, Ioannina, GREECE
[3] Dept. of Economics, University of Ioannina, Ioannina, GREECE

*Corresponding author

kexarcho@cc.uoi.gr, goletsis@cc.uoi.gr, fotiadis@cs.uoi.gr

Abstract. Oral cancer is the predominant neoplasm of the head and neck. Annually, more than 500.000 new cases of oral cancer are reported, worldwide. After the initial treatment of cancer and its complete disappearance, a state called remission, reoccurrence rates still remain quite high and the early identification of such relapses is a matter of great importance. Up to now, several approaches have been proposed for this purpose yielding however, unsatisfactory results. This is mainly attributed to the fragmented nature of these studies which took into account only a limited subset of the factors involved in the development and reoccurrence of oral cancer. In this work we propose a unified and orchestrated approach based on Dynamic Bayesian Networks (DBNs) for the prediction of oral cancer reoccurrence after the disease has reached remission. Several heterogeneous data sources featuring clinical, imaging and genomic information are assembled and analyzed over time, in order to procure new and informative biomarkers which correlate with the progression of the disease and identify early potential relapses (local or metastatic) of the disease.

Keywords: oral cancer, dynamic Bayesian networks, reoccurrence, disease modeling

Introduction

Oral cancer refers to the cancer that arises in the head and neck region, i.e. in any part of the oral cavity or oropharynx. Oral cancer constitutes the eighth most common cancer in the worldwide cancer incidence ranking; more than half million patients are diagnosed with oral squamous-cell carcinoma worldwide every year [1]. Oral cancer is highly related to the sex of the patient, with men facing twice the risk of being diagnosed with oral cancer than women. Research has revealed several risk factors associated with the development of oral cancer. Smoking and excessive consumption

of alcohol, and especially the combination of the two, constitute predominant risk factors for developing oral cancer. Moreover, sun exposure is another important risk factor, particularly for the cancer of the lip [1]. Some studies have also suggested that infection with the human pappilomavirus (HPV) is associated with oral cancer, especially with occurrences in the back of the mouth (oropharynx, base of tongue, tonsillar pillars and crypt, as well as the tonsils themselves) [2].

Cancer cells can spread to other adjacent parts of the neck, the lungs or elsewhere in the body. A common metastasis occurs in the neck lymph nodes through the lymphatic system which helps the cancer cells spread. Although nowadays, the continuous improvements in treatment protocols of cancer have achieved high rates of successful disease disappearance [3], there is a critical stage for the disease evolvement after the treatment called remission; during this stage there is no clinical, laboratory or imaging evidence of the neoplastic mass and the patient is considered cancer free. Nevertheless, even at this point some "invisible" disease particles might still be present leading to a potential spread or metastasis of the disease. Specifically, in terms of oral cancer, locoregional reoccurrence rates after the disease has reached remission have been reported in the range of 25-48%; such high figures can be justified given the deeply infiltrative nature of these tumors, as well as, the significant potential for occult neck metastasis [4].

The reoccurrence rates for oral cancer are quite high and they also suffer from poor prognosis, which can be partly attributed to histologically unfavorable features [4]. Moreover, patients with oral cavity cancer have to deal with the impact of the disease and its treatment on their physical appearance and on the ability to eat and speak, and subsequently with a significant decrease of the quality of life. Hence, early detection of reoccurrence might prove very beneficial [5]. Currently implemented methods aiming to predict oral cancer reoccurrence after the disease has reached remission, have reported quite inadequate results. Although several factors have been associated with the reoccurrence of oral cancer, such as age, site and stage of the primary tumor as well as histological features, they have not been studied altogether in a collective study. Moreover, especially in the molecular basis of the disease, currently available biomarkers are limited in number and efficiency [6, 7]. The efficient combination of the already known ones will greatly benefit the accurate stratification of the patients in terms of staging.

In the general framework of disease prognosis and modeling, several diverse approaches have been proposed in the literature. Most of them involve a prognostic model which implements a risk score depicting the progression of the disease and the general condition of the patient. Based on this score, simple decision rules are used to stratify the patients into several risk categories [8, 9]. More recent approaches utilize advanced machine learning algorithms, such as Artificial Neural Networks (ANNs) or Support Vector Machines (SVMs) which accept as input several variables and provide prediction about the desired outcome. However, most of these approaches use a "black-box" architecture and thus do not provide adequate reasoning about the decision [10, 11]. In addition, it is very cumbersome, if not infeasible to represent properly temporal problems using these algorithms. These issues pose significant limitations for the acceptability of the produced decision systems both by the medical community and the patients. In the case of oral cancer, and cancer in general, the physicians are extremely interested in knowing if, when and why a reoccurrence will

appear. Hence, especially for the problem under consideration (i.e. oral cancer reoccurrence prediction) it is very important to provide sufficient justification about the prediction, but also to introduce the time dimension in the modeling procedure.

In this work, we present an efficient framework in order to systematically study and analyze the factors associated with the reoccurrence of oral cancer, after the remission of the disease. This objective involves the integration of heterogeneous clinical, imaging and genomic data, thus facilitating the multiscale and multilevel modeling of the disease progression over time. Due to the constantly evolving nature of the disease, we employ DBNs, which efficiently cope with temporal causalities, thus, identifying the timing of a potential reoccurrence. Moreover, the intuitive design of DBNs allows for comprehensible decisions coupled with adequate justification. The multitude of gathered data is likely to uncover the evolution and development of the disease during remission, thus assisting the monitoring of patients after treatment, but also contribute towards the accurate stratification of patients in terms of staging. Knowing in advance the progression of the disease, i.e. identifying groups of patients with higher/lower risk of reoccurrence is a key factor towards the determination of the most proper treatment.

Materials and Methods

Clinical scenario

In order to clarify the steps of our study, a clinical scenario is employed which is shown in Figure 1. Initially a patient is diagnosed with cancer through traditional clinical procedures. At this point the physician gathers the required data in order to extract the baseline profile and the patient is treated properly. After the physician's therapeutic intervention, the patient either reaches complete remission or particles of the cancer tissue still remain intact. In the latter case the patients do not qualify for the purposes of our study, whereas from the patients in complete remission, where the cancer is no longer visible, data are further collected, forming the post-treatment profile. Afterwards, and during a two year time span, data are collected from the patient regularly (i.e. scheduled visits are planned for months 1, 3, 6, 9, 12, 15 and 18 after treatment) in order to formulate as a personalized follow-up signature, which is being constantly analyzed. The choice of the follow-up period was determined by the fact that a reoccurrence is most likely to appear in a two year period after the initial treatment. The purpose of this analysis is to stratify the patients in two clusters: i) low risk of disease reoccurrence and ii) high risk of reoccurrence. Hence, we are able to dully identify relapses of the disease and adjust the follow-up treatment accordingly.

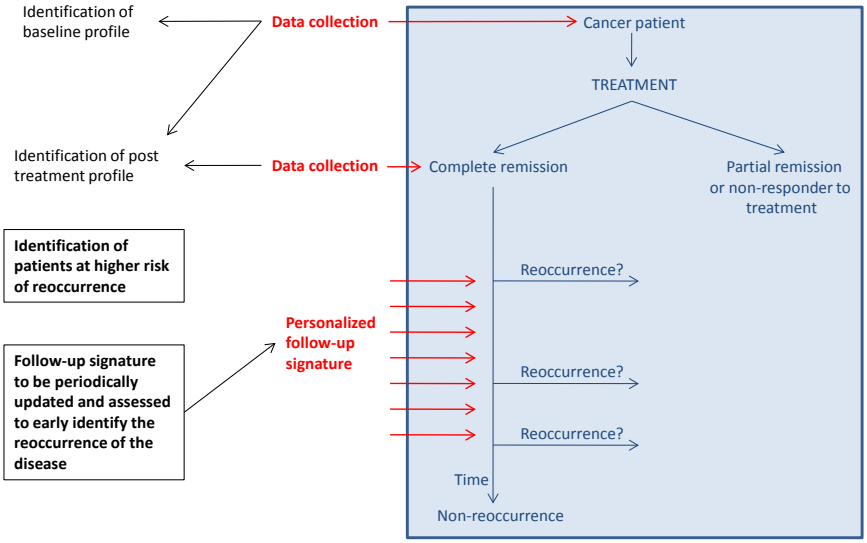

Figure 1: Clinical scenario employed in our study.

Data collection

The progress of the disease in a total of 150 patients with oral squamous cell carcinoma is evaluated during the present study. The cases are collected from two major clinical centers which reside in Italy and Spain. According to available literature 70-80% of these patients are expected to achieve complete remission of the disease after treatment, and an approximate 30-40% of them will develop a reoccurrence of the cancer. Relapses during a two-year time span are marked, as well as the timing of the relapse, and the patients are grouped in two categories, the relapsers and the non-relapsers, which we aim to discriminate by studying and analyzing a multitude of heterogeneous data.

Due to the complex nature of cancer, a major challenge towards its diagnosis and treatment is to formulate a collective approach in order to "frame" every possible aspect. For this purpose we propose a holistic approach which involves the integration and analysis of multiscale and multilevel data. Specifically, clinical, imaging and genomic data are assembled ranging in the scale of dimension and localization. The employment and careful analysis of the above heterogeneous data is likely to reveal the interactions which take place during oral cancer onset and progression. Consequently, the data collected from every patient will comprise the following information:

- Clinical data from health records and standard laboratory markers, histological data from tumor mass specimen

- High throughput genomic data from tumor tissue specimens and circulating cells, profiling gene expression at whole genome level by oligo-RNA microarrays
- Imaging data of the prime tumor mass (and secondary localizations if present)

All these data will be efficiently integrated into a single repository formulating the basis of our study. The data involved in the present study along with the specific techniques employed for their manipulation and analysis are described in detail in the sections that follow.

Clinical

For the diagnosis and monitoring of patients with oral cancer the following types of clinical data are assembled:

- Anamnesis
- Demographics
- Risk factor
- Tumor clinical aspect
- TNM staging
- N characteristics

Anamnesis refers to the detailed medical review of the patient's past health state. Detailed information about the patient's past health problems, general health state, family medical history, oral cancer risk factors and symptoms is gathered in order to establish the diagnosis. Demographic data along with several risk factors are also assembled in order to aid the diagnosis. Next the tumor's clinicopathological stage and developmental phase are evaluated. The most common staging system used for oral cancer is the TNM system. Moreover, several markers have been proven to affect the patient's response to adjuvant and neo-adjuvant treatments [12, 13]. In the present study we compile an extensive list containing all these clinical factors in order to perform a collective study of their relation with oral cancer progression and treatment efficacy. All these data, which comprise the clinical data associated with oral cancer, are thoroughly analyzed for the purposes of the present study.

Genomic

Current advances in the field of genomics have enormously facilitated the thorough analysis of gene expression within cells and tissues. Hence, we are able to extract important information about the interactions and biological pathways which take place during cancer evolution. The framework of the present work employs oligonucleotide and complementary DNA arrays in order to unravel the molecular basis of oral cancer. Nucleic acid arrays have rapidly become a popular investigational tool for cancer biologists, towards the identification of robust genetic biomarkers, thus, shedding considerable light into the complexity of the disease.

Systematic analysis of gene expression data is likely to yield potential tumor markers, or reliable combinations of biomarkers, that can be afterwards used in the daily practice for the diagnosis and monitoring of carcinoma of the head and neck.

Gene expression data come from a feature extraction (FE) file. An FE file is a tab delimited text file comprising of expression values (Log2-ratio data), raw intensity data, background information, metadata regarding the experiment and the scanning settings, gene annotation, etc. A typical FE file is shown in Figure 2.

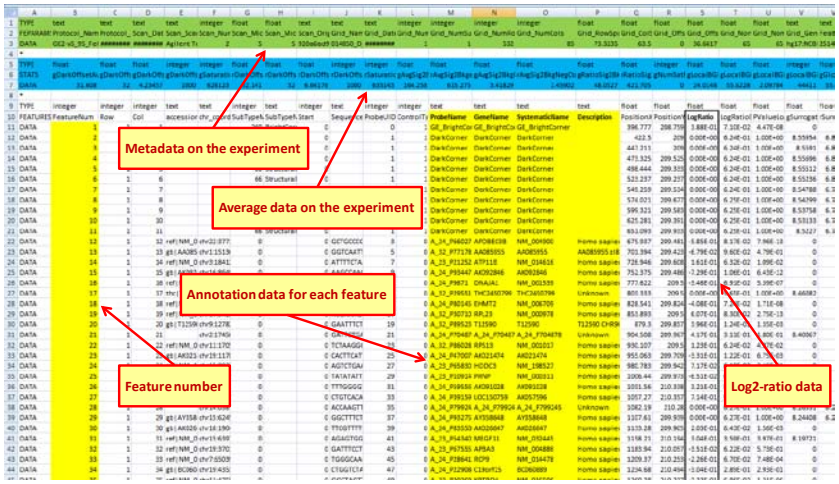

Figure 2: Typical entities extracted from a microarray experiment.

In the present study, all microarray experiments are conducted using the same platform, the same array design and the same FE procedure, in order to minimize the risk of possible sources of variability in the data, other than biological variability.

Especially for genomic data, a preprocessing stage is necessary for enhancing the quality of the data. After obtaining the gene expression data from the microarray experiments the duplicate and control features are eliminated. Control features are negative and positive control elements usually represented by empty features or spots that are hybridized independently from the original sample. Whereas, duplicate features are probes corresponding to a gene or a known internal control sequence which are printed more than once in the array, usually in random positions. They are used to verify the internal consistency of the data and the regional quality of the hybridization. Furthermore, data with high variability, too low signal and genes with a large number of missing values, constituting unreliable expression levels are carefully filtered out.

The overall flowchart for the basic preprocessing of the gene expression data is shown in Figure 3.

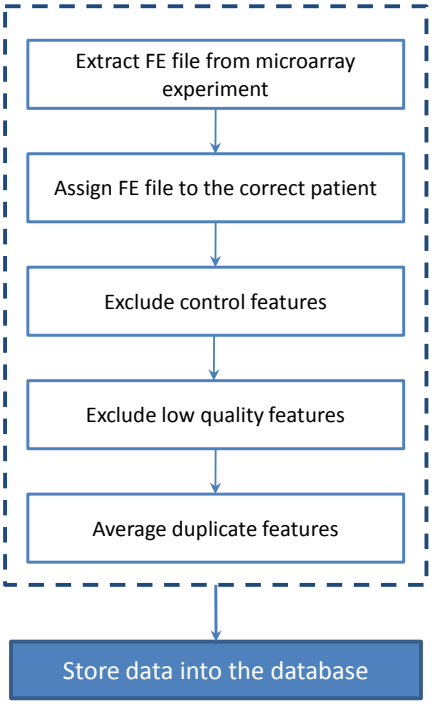

Figure 3: Preprocessing of the gene expression data.

Imaging

Image data from the cancerous tissue can reveal certain significant characteristics of the localization and progress of the disease. The present study employs MRI and CT images. The manipulation of the employed images involves the following main steps, which are also depicted in the flowchart of Figure 4.

- Image preprocessing
- Definition of regions of interest (ROIs)
- Extraction and selection of features
- Classification of the selected ROI

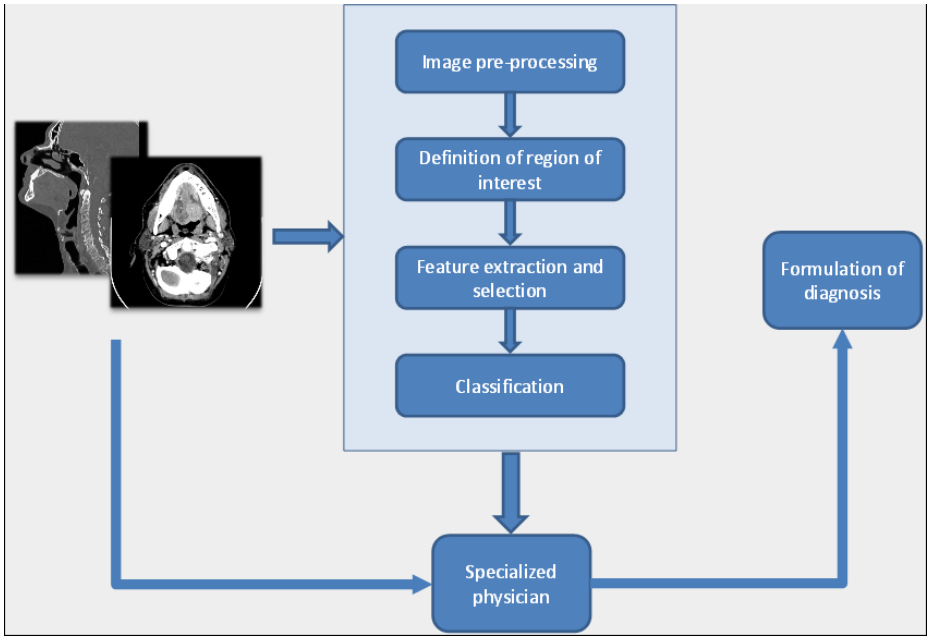

Figure 4: Image data analysis and manipulation.

Initially, the images need to be preprocessed properly in order to improve their quality to facilitate the overall image analysis procedure. The most common types of imaging data contamination are noise and artefacts. Noise causes random distortion in the data, and although several approaches have been proposed in the literature (e.g. application of filters), it is still quite difficult to remove it, due to its random nature. On the other hand, artefacts usually involve more deterministic perturbations of the data, hence it is easier to detect and omit them. These problems can be attributed to several factors such as human error, measuring device limitations, etc. Other types of image preprocessing involve edge enhancement (e.g. unsharping, wavelet transformation), image contrast enhancement (histogram equalization) and image standardization.

In the next step, we detect regions of interest (ROIs), i.e. regions of the preprocessed image bearing enhanced role for our purposes. For the initial approximate definition of some ROIs, a specialized radiologist pinpoints sites of interest, i.e. tumor center, lymph nodes or potential infiltrations. Moreover, automatic methods are also be employed for the detection of ROIs. Active contour models are often employed for automatic definition and tracking of anatomical contours in 2D medical images due to their ability to approximate accurately the random shape of organ boundaries. Seeded region growing is another example of semi-automatic method widely used for the definition of ROIs in medical images.

Afterwards, several features are extracted from the ROIs in order to uniquely characterize the image itself or structures contained in the data. Some of these features represent quantitative measurements with certain physical meaning, that a specialized physician must take into account in order to formulate the diagnosis. However, in some cases features with no apparent physical meaning can be extracted

due to their enhanced discriminative potential. The most common features employed for the analysis of medical images are: pixel based features, texture features, shape features (transformation dependent and transformation independent). Specifically, in the present study, the following features are calculated from each ROI:

- Six (6) features from first order statistics
- Forty eight (48) features from spatial gray-level dependencies matrix
- Twenty (20) features from gray-level differences matrix
- Twelve (12) features from Law's texture energy measurements and
- Three (3) features from fractal dimension measurements

Additional features describing specific properties of the image under consideration are assessed, such as tumor volume, periosteal infiltration, etc. All features extracted during this stage are deposited in a collective repository along with the genomic and clinical data.

Dynamic Bayesian Networks (DBNs)

In the present study we employ DBNs in order to early identify potential relapses of the disease, during the period of remission. As it is described in the clinical scenario, a snapshot of the patient's medical condition is acquired during every predefined follow-up with the doctor. By exploiting the information of history snapshots we aim to model the progression of the disease in the future. The proposed prognostic model is based on DBNs, which are temporal extensions of Bayesian Networks (BNs.) [14]. A BN can be described as $B = (G, P)$ where G is a directed acyclic graph, where the nodes correspond to a set of random variables $\mathbf{X} = \{x_1, x_2, ..., x_N\}$, and P is a joint probability distribution of variables in $\mathbf{X}$, which factorizes as:

$$P(\mathbf{X}) = \prod_{i=1}^{N} P(x_i \mid \pi_G(x_i)) \quad (1)$$

where $\pi_G(x)$ denotes the parents of x in G. A DBN can be defined as a pair $DB = (B_0, B_{trans})$ where B_0 is a BN, defining the prior $P(\mathbf{X_0})$ and B_{trans} is a two-slice temporal BN (2TBN) which defines $P(\mathbf{X_t} \mid \mathbf{X_{t-1}})$. The semantics of a DBN can be defined by "unrolling" the 2TBN until we have T time-slices. The resulting joint distribution is given by:

$$P(\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X}_T) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(x_i^t \mid \pi(x_i^t)) \quad (2)$$

In order to build a model that successfully evaluates the current state or predicts a state in the future (next time slice), we need to train both the structure of the DBN (G_0, G_t) and the parameters of the conditional probability distributions, using both expert knowledge as a prior model and experimental data to get a more accurate posterior model. After the training procedure, we obtain a model as the one shown in Figure 5. By providing some evidence to the model, we are able to compute the

probability of any variable for every time slice (i.e. in any predefined follow-up visit), including of course the probability for reoccurrence.

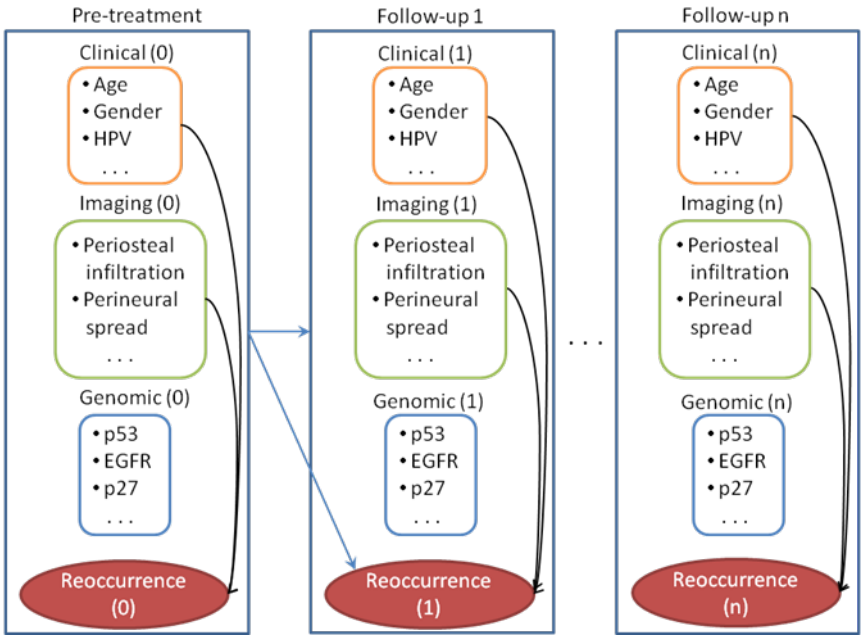

Figure 5: Provisional architecture of the employed DBN model.

For the development of the DBN two implementations have been explored. In the first implementation every source of data is used separately, in order to build a distinct DBN, specifically tailored for a certain type of data. Consequently, three DBNs are developed and their outputs are combined using a meta-classification function (Figure 6(a)). In the second, all sources of data are employed altogether in order to develop a single DBN (Figure 6(b)). However, in both implementations, the contribution and feedback from a specialized doctor, during the DBN construction, is substantial. The two implementations are depicted in Figure 6.

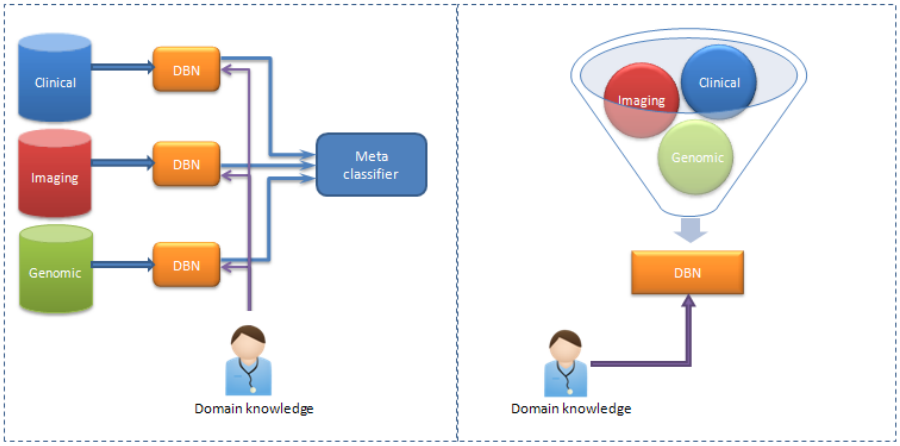

Figure 6: Analysis schemes: (a) multiple DBNs, (b) single DBN.

As this work is currently under development, detailed testing of both implementations depicted in the figure above is needed so as to assess the potential of each one. The assessment will be done using an annotated dataset, covering the two-years follow-up data, which is currently being populated.

Discussion & conclusions

In the present study we propose an advanced framework which implements heterogeneous sources of data towards the prediction of oral cancer reoccurrence in patients that have reached remission. A large amount of clinical, genomic and imaging features are analyzed in order to extract biomarkers that are highly associated with relapses of oral cancer. Thus, we overcome a major limitation of similar studies in the field that employ only a confined subset of features that are associated with oral cancer. Another significant challenge is to capture the disease progression over time. For this purpose we employ DBNs, which are specifically designed to represent temporal causalities. The inclusion of the time dimension is very important as most doctors are interested – even with a rough approximation – in the timing of the reoccurrence. Furthermore, DBNs are able to provide reasoning for the reported decisions, thanks to their transparent architecture. This characteristic is very appealing, if not prerequisite by the medical community. Hence, not only we are able to predict a certain outcome but also to gain insight about the rationale of every decision. In overall, the currently proposed framework contributes significantly towards the monitoring of oral cancer evolvement since it can answer if, when and why a reoccurrence might appear.

Acknowledgements

This work is part funded by the European Commission NeoMark project (FP7-ICT-2007-224483) – ICT enabled prediction of cancer reoccurrence.

References

1. Haddad, R.I., Shin, D.M.: Recent advances in head and neck cancer. The New England journal of medicine **359** (2008) 1143-1154
2. Mork, J., Lie, A.K., Glattre, E., Hallmans, G., Jellum, E., Koskela, P., Moller, B., Pukkala, E., Schiller, J.T., Youngman, L., Lehtinen, M., Dillner, J.: Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck. The New England journal of medicine **344** (2001) 1125-1131
3. Forastiere, A., Weber, R., Ang, K.: Treatment of head and neck cancer. The New England journal of medicine **358** (2008) 1076; author reply 1077-1078
4. Godden, D.R., Ribeiro, N.F., Hassanein, K., Langton, S.G.: Recurrent neck disease in oral cancer. J Oral Maxillofac Surg **60** (2002) 748-753; discussion753-745
5. Sciubba, J.J.: Oral cancer. The importance of early diagnosis and treatment. American journal of clinical dermatology **2** (2001) 239-251
6. D'Silva, N.J., Ward, B.B.: Tissue biomarkers for diagnosis & management of oral squamous cell carcinoma. The Alpha omegan **100** (2007) 182-189
7. Lippman, S.M., Hong, W.K.: Molecular markers of the risk of oral cancer. The New England journal of medicine **344** (2001) 1323-1326
8. Knaus, W.A., Wagner, D.P., Draper, E.A., Zimmerman, J.E., Bergner, M., Bastos, P.G., Sirio, C.A., Murphy, D.J., Lotring, T., Damiano, A., et al.: The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest **100** (1991) 1619-1636
9. Le Gall, J.R., Lemeshow, S., Saulnier, F.: A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. Jama **270** (1993) 2957-2963
10. Cruz, J.A., Wishart, D.S.: Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics **2** (2006) 59-78
11. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine **34** (2005) 113-127
12. Woolgar, J.A., Rogers, S., West, C.R., Errington, R.D., Brown, J.S., Vaughan, E.D.: Survival and patterns of recurrence in 200 oral cancer patients treated by radical surgery and neck dissection. Oral oncology **35** (1999) 257-265
13. Woolgar, J.A., Scott, J., Vaughan, E.D., Brown, J.S., West, C.R., Rogers, S.: Survival, metastasis and recurrence of oral cancer in relation to pathological features. Annals of the Royal College of Surgeons of England **77** (1995) 325-331
14. Murphy, K.P.: Dynamic Bayesian Networks: Representation, Inference and Learning. UNIVERSITY OF CALIFORNIA (2002)