



Structuring European Biomedical Informatics to Support Individualised Healthcare

IST-507585

<http://www.infobiomed.org>

**State of the Art on Data
Biomedical Informatics in Chronic
Infectious and Inflammatory Disease
Research: Periodontitis as a Case Study**

WP6.3 – Genomics and chronic inflammation

**Internal Deliverable
Final**

Authors (affiliation): ACTA, AVEIRO, CUSTODIX, ISCHII, LEICESTER, UPM

Lead participant: ACTA

Date: 22/12/2005

Type: Internal Deliverable

Dissemination level: Public

INFOBIOMED - Pilot 6.3
STATE OF THE ART REPORT (SOA)
**Biomedical Informatics in Chronic Infectious
and Inflammatory Disease Research:
Periodontitis as a Case Study**

Partners:

ACTA
AVEIRO
CUSTODIX
ISCIH
LEICESTER
UPM

Please address correspondence to:

Dr. Bruno G. Loos
Department of Periodontology
Academic Center for Dentistry Amsterdam
Louwesweg 1
1066 EA Amsterdam
The Netherlands
e-mail: B.G.Loos@acta.nl
Fax: + 31 20 518 8512

TABLE OF CONTENTS

- 1 Background
 - 1.1 The study of complex diseases
 - 1.2 Periodontitis as a model for a complex disease
 - 1.3 Current research objectives
- 2 Sources of information for studies into periodontitis
 - 2.1 Disease phenotype
 - 2.2 Genetics
 - 2.3 Infection
 - 2.4 Environmental (behavioral) factors
 - 2.5 Intermediate phenotypes
- 3 Challenges for Biomedical Informatics in the study of periodontitis
 - 3.1 Informatics in relation to complex diseases
 - 3.2 Informatics needs in periodontitis
 - 3.2.1 Data collection
 - 3.2.2 Annotating and standardizing data
 - 3.2.3 Representing and assessing the phenotype
 - 3.2.4 Ensuring privacy
 - 3.2.5 Integrating and standardizing with external sources
 - 3.2.6 Mining and visualizing data
- 4 Relevance to the study of chronic inflammatory diseases
- 5 Conclusions

1. Background

1.1 The study of complex diseases

Common diseases, such as cardiovascular disease, cancer, metabolic imbalances, and chronic inflammatory illnesses impose a major drain on society, both in terms of financial and human costs. They are thus priority targets for current biomedical research. Chronic inflammation is of particular interest, since beyond obvious inflammatory states, it may well be a contributory factor in a range of common disorders (e.g., Crohn's Disease, Alzheimer's disease and atherosclerosis). As for all common diseases, inflammatory disorders are likely to have an etiology that is highly complex, entailing many interacting genetic and environmental factors. Hence the name 'complex disease' is often used when referring to common disorders. This complexity has been an obstacle to effective research in this field. However, as 'genomics' technologies now become commonplace and increasingly effective, there is hope that real progress can be made – not least for chronic inflammatory disease.

Typical human chronic inflammatory diseases include multiple sclerosis, rheumatoid arthritis, atherosclerosis, Crohn's disease, and periodontitis. Chronic inflammatory diseases have a complex pathogenesis and have a multifactorial etiology, involving complex interactions between multiple genetic loci, infectious agents and environmental (behavioral) factors such as diet, smoking habits and physical exercise (Tabor et al. 2002). The role of infectious agents (bacteria and viruses) in particular, is gaining increased interest and recognition (Kuipers et al. 2003, Liuba 2004, Mukamal et al. 2004). The general paradigm is that certain individuals are genetically more susceptible than others to the environmental risk factors (Tabor et al. 2002), and it is these persons that are more likely to succumb to the illness. This innate variability then also explains why patients vary in age of onset, severity and response to medical treatment. Furthermore, complexity arises since very different pathological processes could lead to the same final clinical presentation. Since multiple genes contribute to the susceptibility and to the course of the disease, differences will also exist between different geographical populations that have different spectra of genome variations. This phenomenon, however, is superimposed upon the very different lifestyles, environments, and risks of infection that exist across the world.

1.2 Periodontitis as a model for a complex disease

Periodontitis is a chronic inflammatory disease of the supporting tissues of the teeth. If left untreated, teeth may show exposed root surfaces, in conjunction with red, swollen gums that easily bleed. Dental radiographs reveal periodontal (alveolar) bone loss around the teeth that is due to the inflammation process (Figure 1); teeth will become mobile and migrate, and will eventually exfoliate. Patients with periodontitis experience problems with chewing due to tooth mobility and loss of teeth; they have bad breath and suffer from important subjective and objective esthetic problems. Dental practitioners provide labor-intensive diagnostic and treatment sessions to periodontitis patients, including periodontal surgery.

Recent data suggest that periodontitis is associated with increased risk for cardiovascular diseases (Beck et al. 1999), possibly through the elevation of the acute phase reactant C-reactive protein (CRP) or other systemic markers of inflammation (Loos 2005). The systemic reactions to periodontitis may come about through systemic dissemination of oral bacteria. There are strong indications that the inflamed and ulcerated pocket epithelium forms an easy port of entry for oral microorganisms. Short moments of bacteremia occur most likely several times a day. Like any other inflammatory condition, untreated chronic periodontitis may pose a risk for the overall health of the subject (Park 2002).

Periodontitis has a relative high prevalence in the population. It has been estimated that about 10% of the total adult population and about 30% of individuals over the age of 50 years, suffer from severe periodontitis (Brown et al. 1990, Gjermo 1998). When destructive periodontal disease progresses at a relative slow pace and is diagnosed during middle age, we recognize this form as a chronic, adult form of periodontitis. However, in some individuals the disease manifests itself at adolescent or post-adolescent age in a rapidly progressive manner; in this situation we like to speak about early onset periodontitis or aggressive periodontitis (Van der Velden 2000). Unfortunately however, there is no global consensus on a diagnostic system and classification scheme for periodontitis (Van der Velden 2005). Interestingly, a recent study indicated that three experienced clinicians, adhering strictly to the proposed general classification scheme (Armitage 1999), diagnosed 25% of cases differently (Picolos et al. 2005). Moreover and even more important, it is highly likely that several forms of periodontitis exist, which ultimately present with similar clinical features, i.e. gingival inflammation, deepened bleeding pockets, loss of periodontal attachment and loss of alveolar bone, resulting in loose and non-functional teeth.

Thus it is accepted that the etiology of periodontitis is multifactorial (Page et al. 1997) involving:

1) *Infectious component.* A large variety of oral bacteria are able to colonize the subgingival region, i.e. the area directly around the teeth below the gum line. These bacteria form dental plaque, which is attached to the root surfaces of the teeth. It is recognized that in the periodontal lesion (the pocket), bacteria are organized in a complex microbial biofilm, which consists predominantly of strict anaerobic, mainly Gram-negative bacteria (Chen 2001). Of the several hundred oral bacterial species known, a limited number of species are recognized as periodontal pathogens and they have been identified as important markers of progressive disease. These include: *Porphyromonas gingivalis*, *Prevotella intermedia*, *Tannerella forsythensis*, *Fusobacterium nucleatum*, *Actinobacillus actinomycetemcomitans* and the spirochaetal species *Treponema denticola* (van Winkelhoff et al. 2002). The complexity of the subgingival biofilm as well as the interplay of various species associated with periodontitis is also exemplified by the proposal that infective and protective bacterial species are organized in microbial complexes (Socransky et al. 1998, Socransky et al. 2002). In this way a red complex includes the most infective species *P. gingivalis*, *T. denticola* and *T. forsythensis*, while on the other end of the spectrum the yellow complex is consisting of Streptococcal species (Table 1).

It is important to note that not all patients with periodontitis are infected with the same and all the periodontal pathogens. The microbiological component of periodontitis varies considerably among patients, which makes periodontitis also complex in microbial sense. Moreover, it has been suggested that not all bacteria associated with periodontitis are normal inhabitants of the oral cavity (Griffen et al. 1998, van Winkelhoff et al. 2002). *A. actinomycetemcomitans* and *P. gingivalis* show characteristics of exogenous pathogens, a view that is based on their low prevalence in periodontal health (Slots & Ting 2002).

2) *Genetic susceptibility.* Genetic susceptibility for chronic periodontitis is deduced from family studies and studies in twins. It is recognized that frequently siblings of patients with early onset aggressive periodontitis also suffer from periodontitis (Boughman et al. 1992). On the basis of a study in family units with 3 or more siblings, also the adult form of chronic periodontitis was shown to have a genetic background (Van der Velden et al. 1993). From twin studies it has been estimated that 38% to 82% of population variance in periodontal disease expression, may be

attributed to genetic factors (Michalowicz et al. 1991). Chronic adult periodontitis was estimated to have 50% heritability, which was unaltered following adjustments for behavioral variables including smoking (Michalowicz et al. 2000).

The search for genetic markers and candidate disease-modifying genes in periodontitis has recently been receiving considerable attention. In particular, single nucleotide polymorphisms (SNP's) in genes encoding molecules of the host defense system have been targeted (immunogenetics) (Table 2) (Loos et al. 2005). Analogous to other complex inflammatory diseases, it is clear that periodontitis is a polygenic disorder. In this way modifying disease genes have been identified in the interleukin-1 gene cluster (Kornman et al. 1997, Laine et al. 2001) and Fcγ receptor loci (Loos et al. 2003, Meisel et al. 2001, Yamamoto et al. 2004). Moreover, SNP's in the interleukin-10 and vitamin-D genes have been associated with periodontitis (Berglundh et al. 2003, de Brito Junior et al. 2004, Gonzales et al. 2002, Hennig et al. 1999, Scarel-Caminaga et al. 2004, Sun et al. 2002). In addition, also focus of attention is directed at genes that encode proteins playing a role in the innate immunity, such as CD14 and Toll-like receptors (TLR) (Loos et al. 2005).

Periodontitis can also be observed as a component of several single gene syndromes (Kinane & Hart 2003). These disorders are characterized either by immune or structural deficiencies; of these syndromic disorders, the Papillon Lefèvre syndrome (PLS) is relatively unique, in that periodontitis forms a significant component of the disease and indeed, is one of two defining clinical features. Mutations in the cathepsin C gene (*CTSC*) have been identified as causal for PLS, which includes prepubertal periodontitis (PP) (Hart et al. 1999, Toomes et al. 1999). Interestingly, some *CTSC* mutations are causal for PP without PLS (Hart et al. 2000). No relationship has been demonstrated between *CTSC* mutations and other forms of periodontitis (Hewitt et al. 2004).

3) *Environmental factors*. Smoking is currently accepted as the most important environmental risk factor in periodontitis (Bergström 1989, Haber et al. 1993, Papapanou & Lindhe 2003). Periodontitis susceptible subjects suffer from a more progressive form of periodontitis and smokers have more severe periodontal breakdown than nonsmoker patients (Calsina et al. 2002, Xu et al. 2002). Furthermore, smokers with periodontitis show a less favorable response to periodontal treatment, both for non-surgical and surgical approaches (Grossi et al. 1996).

The role of smoking in the disease process has recently been reviewed (Palmer et al. 2005), but is actually still unknown. Some papers have suggested that there may be differences in the subgingival microflora between smokers and non-smokers with periodontitis (Van Winkelhoff et al. 2001), but others did not find this relation (Bostrom et al. 2001). Smoking may also hamper the host resistance and immunological functions (Palmer et al. 2005). Some reports have suggested a reduced phagocytosis, altered T cell function and lack of immunoglobulin production in smokers with periodontitis compared to non-smoker patients (Graswinckel et al. 2004, Loos et al. 2004, MacFarlane et al. 1992). Another important hypothesis for smoking as risk factor is that it may reduce local blood supply in the periodontal tissues (Bouclin et al. 1997).

Other factors that have been proposed as environmental risk factors for periodontitis include diet and stress (Al-Zahrani et al. 2005, Breivik et al. 1996, Chapple 1997, Loos et al. 1998, Staudte et al. 2005). However these factors are not yet firmly established; more research is needed into their role in the etiology and pathophysiology of periodontitis.

1.3 Current research objectives

In spite of the above summary of the current paradigm of periodontitis, little is known on the pathophysiology. The disease develops as a consequence of the interrelation of genetic predisposition, genetic modulation of the immune response and the role of bacteria, smoking and other behavioral factors such as diet and stress. It is a challenge to increase our understanding of the pathogenesis of periodontal diseases. Periodontitis is deemed to be an excellent model for BMI, because of its multifactorial etiology, *i.e.* infectious component, genetics, environmental factors. Data can be easily obtained due to the relative high disease prevalence, plus the fact that periodontitis is not a life threatening disease and that no invasive procedures are necessary to obtain biological samples. Furthermore patients are often very willing to collaborate. Diseased and healthy tissues, genomic DNA and access to the history of infections and other relevant data through the patient records, are all accessible. Integration of various data sources will give new insights in the pathophysiology, will enable us to make new disease classification schemes and ultimately will result in risk profiling and screening. To meet these objectives, the disciplines periodontology, MI and BI have to be integrated.

Recently the Network of Excellence “INFOBIOMED” has been founded. The goals of BMI within the INFOBIOMED consortium for the improvement of human health are, among others, the

integration of genetic and clinical information with preventive, diagnostic and therapeutic purposes (Sanz et al. 2004). Within the consortium of INFOBIOMED, the complex disease periodontitis will serve as a case study and model of integration of BI and MI. Figure 2 summarizes the factors involved in this disease and their interrelations providing a structure for the complexity of its study.

2. Sources of information for studies into periodontitis

Necessary data on disease manifestations (phenotype), patient histories, environmental factors and behavioral aspects, genotyping and microbiological results, diagnosis and treatment procedures are stored in different databases, in different formats and most of the times at different locations.

2.1 Disease phenotype

The classification system of periodontitis is still a matter of debate. Two forms of the disease, *i.e.* chronic periodontitis and aggressive periodontitis, have been proposed (Armitage 1999). However this classification is clearly unsatisfactory, because within these groups differences in extent, severity and age of onset vary. Therefore a new classification scheme has been proposed (Van der Velden 2000, 2005). This scheme encompasses several features including extent and severity as well as age of onset.

Periodontitis is clinically defined by the presence of deepened pockets (>4 mm) concomitant with loss of probing attachment. The pockets and loss of attachment are measured with a periodontal probe mostly at 6 sites around each natural tooth. These clinical measurements are relatively simple to perform but must be done by a trained and calibrated dentist or dental hygienist. The amount of deepened pockets and mean clinical attachment loss are indicative of the extent of periodontal destruction. The clinical measurements are sometimes painful for the patient. An alternative method to determine periodontal destruction is by scoring periodontal bone loss on dental radiographs. To this end, from 10 up to 18 dental radiographs are made, each providing a good image of one or two complete teeth. Traditionally, dental radiographs are made on photographic film. Recently dental radiographs are made digitally and are stored as *.tif* files. Nevertheless, it is possible to digitize the traditional radiographs by DICOM protocol.

To determine the amount of periodontal destruction on radiographs around each tooth, the distance is measured between the cemento-enamel junction (CEJ) and periodontal bone level; this can be done in mm's, or in % of the total root length. Measurements are performed manually identifying 3 landmarks: the CEJ, the periodontal bone level and the apex (Figure 1). The dependency on the clinician or researcher to set the landmarks is a disadvantage of the method. Another disadvantage is the labor intensive work for the clinician or researcher to identify for each tooth, both on the mesial and distal aspects, the three landmarks and then to calculate the percentage of bone support lost.

Currently, exchange of digitized radiographs between research centers, is not routinely done. There is no standard applied; we recommend storing radiographic data according to the DICOM standard. An advantage for this standard is that patient demography data can be coupled to the images and the chances for mislabeling are diminished. Radiographic data in DICOM can be accessed by any computer system, which may be an important advantage in the studies of complex diseases, where multiple databases exist.

Therefore it is to be investigated how existing digitized radiographs and radiographs in .tif application, can be converted to the DICOM standard for subsequent automatic analysis for periodontal destruction.

2.2 Genetics

It will be challenging to understand the genetic basis of complex diseases, and there it is needed to constantly (re-)evaluate the relevant issues. An analogy with a jigsaw puzzle has been made (Brookes 2001), in that the challenge may entail too many pieces (genes), many of which may be of minor contribution (modifying genes). The candidate genes may even be from different jigsaws (different disease etiologies), and there are many ways the jigsaw can be partially though incorrectly assembled. False positive findings seem to be abundant, due to a number of causes that will be difficult to isolate or overcome. It is estimated that meaningful results for the candidate gene approach may only be obtained with 1000's of patients, since most associations refer to small odds ratio's (range 1.1 - 1.50) (Clayton & McKeigue 2001, Colhoun et al. 2003, Ioannidis 2003, Ioannidis et al. 2003, Tabor et al. 2002). But perhaps the main problem is that with complex diseases we cannot know in advance how much actual complexity there is to be dealt with, nor whether available tools and clinical materials come anywhere close to being able to manage the

task. To help, we can employ both MI and BI, in ways that are complementary. That is, to aid the physician who wants to determine risk of a certain disease, BI can bring guiding knowledge to the MI world, whilst to help the researcher understand the pathophysiology, MI provides the disease phenotype and related data for BI analysis.

Regarding periodontitis, the genetic information needed for MI and BI activities is spread over different databases of simple structure (SAS, SPSS, Excel). It is thus problematic to attain an up-to-date and comprehensive view of what is known and what is being tested, without manually clicking through numerous different articles and web pages.

Need from the consortium: Improved systems for database integration and data mining. This could entail enforcing greater data linking, simplifying tools for integrating data views across databases (e.g., Distributed Annotation System protocol), or promoting the construction of centralized data warehouses. In either case, the field would be helped substantially by the development and adoption of standards for data representation and exchange, in particular with regards to phenotype data.

2.3 Infection

Bacteria play an essential role in the pathogenesis of periodontitis. Although over 500 different oral species have been proposed to exist, a limited number of bacteria have been associated with periodontitis (Socransky et al. 1998, Socransky et al. 2002, van Winkelhoff et al. 2002). The prevalence of suspected periodontal pathogens may be an indication for an active periodontal disease process. The dentist combines the bacterial situation with other systemic and environmental factors, like smoking habits, to determine diagnosis, prognosis and to evaluate treatment outcome and future disease development.

For understanding the pathophysiology of periodontitis, for possible new classification schemes and to integrate the infectious component with the genetic susceptibility, the data on the bacterial infection needs to be standardized. Standard diagnosis of the total bacterial load and 7 suspected pathogens (Taxonomy NCBI) is performed. The species include *A. actinomycetemcomitans*, *P. gingivalis*, *P. intermedia*, *T. forsythensis*, *F. nucleatum*, *M. micros* and *E. corrodens*. These data are currently in written reports and need to be coupled to the genetics and environmental data sets.

Need from the consortium: Generate data entry tools to be used by the laboratory personnel, for the generated BI which can directly be integrated with other BMI on the same patient.

2.4 Environmental (behavioral) factors

Environmental aspects involved in periodontal diseases are multiple and variable and involves smoking, stress, diet (vitamins), socio-economic status (SES), educational level, medical status, living area, stress, oral hygiene and others. The environmental factors related to periodontitis patients are normally established by questionnaires. The disadvantage of questionnaires is the fact that the interpretation of the questions depends on both the physician and patient. From a BMI standpoint, the difficulty of questionnaires is “free text”. Both statistics and databases cannot handle properly these kinds of data.

Need from the consortium: Generate data entry tools to be used by the patient, which can directly be integrated with other BMI on the same patient.

2.5 Intermediate phenotypes

The parameters age, gender, ethnic background, weight, height, body mass index, blood pressure and biochemical variables (triglycerides, cholesterol, glucose, insulin, C-reactive protein) briefly summarize the great variety of reference terms that can be encountered in different studies into periodontitis. These variables are defined as intermediate phenotypic data and they are important factors to be included in the overall diagnosis of the disease and may be (strongly) related to disease expression. Moreover family history, disease expression in siblings and parents and/or children, needs to be considered. The generated data are mostly measured numbers and reference values. Obviously if a genetic polymorphism is associated with an intermediate phenotype, for example the circulating product of this gene, and both the polymorphism and the intermediate phenotype are associated with the disease endpoint in a coherent way (not in opposite directions), the results may be said to be internally consistent. This is not a proof of causality, but it is better than an isolated association between the polymorphism and the disease. Actually, internal consistency is one of the criteria used to support causality in epidemiological studies. This is why as much information as possible should be collected *a priori* that could be used to assess consistency and intermediate phenotypes.

Intermediate Phenotype (Physiome or Phenome)

- Physiology Visualization (PET, fMRI,...)
- Physiological Assay (BME instruments)
- Analytical Assay (chemical, proteome)

In fact, the use of such an intermediate phenotype may prove crucial in elucidating the physiologic mechanism(s) by which an identified relevant genetic variant imparts an effect.

Need from the consortium: Generate data entry tools for intermediate phenotypic data to be used by the dentist/dental researcher which can directly be integrated with other BMI on the same patient.

3 Challenges for Biomedical Informatics in the study of periodontitis

3.1. Informatics in relation to complex diseases

Informatics is increasingly pertinent in medicine. First, a discipline called Medical Informatics (MI) has been traditionally involved in the research and development of informatics-based methods and tools for medical and epidemiological research and patient care and management. Bio-Informatics (BI) on the other hand, is a discipline developed around the Human Genome Project, which handles genomic, proteomic and any other biological research data. MI and BI can play an important role in our understanding of complex diseases. Both MI and BI will contribute to gain new insights in complex diseases, will help to better understand the pathophysiology of inflammatory processes, and will help to develop new classification schemes, design new treatment strategies and ultimately preventive and pro-active measures.

However, MI and BI are disciplines that up to now have followed separate developments with few contacts and synergies between them. It is the elucidation of the human genome that has evidenced the need and the possibilities for a strong synergy between the two. It is no longer sufficient to separately undertake and computationally support clinical practice, clinical research, and classical epidemiological and genomic research, to meaningfully advance the so-called genomic medicine. Instead a new integrative approach is required. The integration and exploitation of all the data and information generated at all levels by these disciplines require a new synergistic approach.

This approach will enable a two-way dialogue that will combine data, methods, technologies, tools, applications and views. Biomedical Informatics (BMI) is the emerging discipline that aims to put these two worlds together so that the discovery and creation of novel diagnostic and therapeutic methods is fostered (Martin-Sanchez et al. 2004).

The mission of BMI is to provide the technical and scientific infrastructure and knowledge to allow evidence-based, individualised healthcare using relevant sources of information. These sources include the traditional information as currently maintained in the health record, as well as new genetic and other biological information. Aiming at a change from late stage diagnosis towards early detection or even prediction and prevention of disease, BMI bears the potential to improve health and quality of life of the individual, and reduce overall costs of healthcare systems.

3.2 Informatics needs in periodontitis

Periodontitis is a complex disease that requires the integration and analysis of multiple sources of data and it thus offers a remarkable set of challenges to BMI. These include infrastructural challenges such as the creation of data models and databases for storing these data, the integration of these data with external databases, the extraction of information from natural language text, and the protection of databases with sensitive information (Sujansky 2001). If these challenges are met, then the resulting applications and tools will have clear and specific impact in the periodontal research area as well as in clinical practice. To fully realize the potential of the emerging interdisciplinary field of BMI, we believe that it is necessary to move toward the establishment of common infrastructure for exchanging periodontitis data. The creation of an internationally compatible informatics platform for exchanging these data will enhance the impact of the individual datasets and provide the scientific community with easy access to integrated data in a structured standard format, facilitating data exchange, and comparison and data analysis. The objective is to develop a BMI tool that will provide support for clinical decision-making, for the integration of clinical-genomic data of patients in a repository, for determining patient risk profile and ultimately for the reclassification of the disease based on the new genetic information collected and that is being included in various research centers.

Even though the focus of this pilot mainly has initially application in the classification and biology of the disease, it will eventually be brought into daily clinical practice. The demands of researchers in the field of periodontitis clearly show the needs for the development of a BMI

structure and protocol. The main challenges for biomedical informatics within pilot 6.3 fall into next areas:

1. Data collection
2. Annotating and standardizing data
3. Representing and assessing the phenotype and life habits data
4. Ensuring privacy
5. Integrating and standardizing with external sources
6. Mining and visualization of data

3.2.1 Data Collection

Supporting the whole structure we have the enabling technologies oriented towards facilitating a solution for a periodontitis data warehouse (PDW). The PDW will include parameters that are collected in daily clinical practice that respond to five distinct levels: genetic, environmental (behavioral), infectious, disease phenotype and intermediate phenotype (individual risk factors) (see Table 3).

Genotype

To understand the genetic basis of periodontitis molecular techniques will be applied in genetics laboratories to advance in the knowledge of SNPs, haplotypes, mutations or genes involved in the development of the disease. Background and basic information needed for these activities is stored over many different databases like dbSNP, OMIN, HGNC, Entrez Gene and GO; this involves a major complexity at the moment of integrating the generated information and also to update this information. These problems can be solved constructing a centralized data warehouse, like PDW, that promotes the development of standards for data representation and exchange, and which is linked to the various external databases.

Environmental (behavior) factors

The environmental factors related to periodontitis like smoking, diet (vitamins), oral hygiene habits, medications, and demographic data like socio-economic status, educational level, medical status, living area, are normally gathered by questionnaires, clinical interviews or statistics that can be consulted in hospital/clinic and epidemiological databases.

Infectious agent

The role of endogenous and/or exogenous bacterial pathogens in the pathogenesis of periodontitis necessitates the collection of microbiological data from cultures, PCR- or checkerboard generated data, as well as antibiograms, to identify the (clusters of) microorganisms involved. There is a need for the use of a standard bacterial nomenclature and also it is necessary consult related information and background from existing databases like DSMZ or NCBI Taxonomy. The main matter associated with the understanding of the role of the microorganisms in the disease is the need of a standardization, which unifies genetics and environmental data sets.

Disease phenotype

The determination of the phenotypic manifestation of periodontitis is based on the examination of the mouth and teeth, the deposits of plaque and calculus, a gingival bleeding index, probing pocket depth and attachment loss measurements, as well as the measurements of loss of supporting bone; this information may be stored in Clinical Patient records (CPR). Bone loss from patients will be determined through X-rays that are recommended to be digitized and stored by DICOM protocol. The images in digital form necessitate the development of a digital image management system. Such systems, often referred to as Picture Archiving and Communication Systems (PACS), are emerging in clinical and radiological environments. The PACS designs are geared toward a centralized system where images and data are stored in a central archive and further distributed upon request to peripheral workstations. With the development of network facilities and inter-computer communication systems, it becomes possible to design distributed archives where images and data can be stored at different locations of a network and still be accessible in any other part of that network.

Intermediate phenotypes

Patient clinical, biochemical or physiological data like age and gender, ethnic background, weight/length and body mass index, blood pressure, are extracted from the hospital information systems, personal interviews, clinical examinations and analytical laboratory tests. The generated data can be integrated in CPR or in the laboratory information system to be consulted by the dentist/dental researcher.

3.2.2 Annotating and standardizing data

Research data, including protocols, primary data collected from the phenotype and intermediate phenotype, information coming from external sources and procedures of reduction and

analysis, are the essential components of scientific progress. All the mentioned information is being stored in a Data Warehouse. Attention should be given to annotating and documenting computerized information to facilitate detailed review of data that should be treated comparably. The use of controlled vocabularies and nomenclature is needed to enable database queries and automated data analysis for representing and exchanging data. Controlled vocabularies and nomenclature is also designed to facilitate the retrieval and integration of information from many information sources, including clinical records or biological knowledge bases. An example of the latter is DSMZ, which gives a standard annotation depending on the place within the phylogenetic framework. Another example is HGNC that provides for each gene only one given and approved gene symbol to facilitate data retrieval. Finally the use of GO can help to annotate and standardize data; it provides a controlled vocabulary to describe gene and gene product attributes in any organism.

3.2.3 Representing and assessing the phenotype and life habits data

Aveiro, one of the technological partners, has developed an ad-hoc tool to collect and analyze information from dental x-rays that will eventually be connected to the database at ACTA. Although there is no standard applied, it is recommended storing radiographic data according to the DICOM standard. An advantage for this standard is that patient demography data can be coupled to the images and the chances for mislabeling are diminished. Radiographic data in DICOM can be accessed by any computer system, which may be an important advantage in the studies of complex diseases, where multiple databases exist.

3.2.4 Ensuring privacy

For the studies into periodontitis, patient data will be generated. These include disease and intermediate phenotype, genetic data, data on the infectious component and environmental factors. From different sides an integrated data set must be generated and is to be used by dentists and dental researchers, as well as researchers in the field of informatics. To ensure the privacy of the patient an accepted coding system needs to be applied.

The growing need of managing both clinical and genetic data raises important privacy protection challenges. Privacy includes the right of individuals and organizations to determine for themselves when, how and to what extent information about them is communicated to others. The

differences between genetic information and other medical information can be summarized as follows:

- Genetic data not only concern individuals, but also their relatives, thus people who have not been tested directly;
- Medical data deal with past and current health statuses of persons, whereas genetic testing can also give indications about future health or disease conditions;
- Personal genetic profiles can directly be derived from tissue samples;
- An individual person's genotype is almost unique and stable;
- The potential information content of genetic data is not known.

Several basic approaches to safeguarding data confidentiality have been identified in the past. The first approach focuses on the creators and maintainers of the information, prohibiting them from disclosing the information to inappropriate parties. An alternative approach focuses on the use of so-called Privacy-Enhancing-Techniques (PETs). Privacy enhancing solutions range from very simple to complex technical and non-technical methods and measures. Examples of such techniques are (non exhaustive list):

- Hard de-identification at the source side by the owner of the data;
- Various types of de-identification (anonymization and/or pseudonymization) which may be reversible or irreversible, conducted with or without the help of a Trusted Third Party (TTP), in batch or in interactive mode, etc.;
- Controlled database dilution/perturbation, which consists in injecting fake data in a controlled way;
- Data (flow) segmentation;
- Diagnostic (preventive) privacy protection gauging of databases in order to calculate direct and indirect (re-) identification risks;
- Privacy enhancing software agents that interact intelligently to continuously check and ensure the level of privacy protection in a database.

The study on the complex disease periodontitis is requiring immediate source interactions, typically remote database access. Information is interactively obtained from, or delivered to the user at the source. Today such database access is often implemented with web browser technology. In the interactive (reversible) pseudonymization model from Custodix (De Moor et al. 2003), a (transparent) intermediary privacy protection engine is put between the users and the database web

server. Such solution renders the interposition of a privacy protection service largely transparent to both the user and the database web server. Nothing is changed from the perspective of the user of the on-line database application, while non authorized end-users such as researchers, can still get access to the data in the pseudonymized database. The implementation of such PETs may possibly provoke, for a number of applications, a shift in paradigm, namely from “Privacy through Security” to “Privacy through Privacy”.

3.2.5 Integrating and standardizing with external sources

Some data exchange standards have been developed like the eXtensible Markup Language (XML), which is a standard syntax for specifying how text data can be labeled within the file and how data are presented.

3.2.6 Mining and visualization data

The data warehouse has large quantities of heterogeneous data collected from diverse sources mentioned above and this information was stored by specific categories so it can be more easily retrieved, interpreted, and sorted by users. But this vast collection of raw data is not in themselves useful. To be meaningful, data must be analyzed and converted into information, or even better, into knowledge. Data mining tools and knowledge discovery techniques applied to databases, are very promising approaches to help answer research questions and provide information in meaningful ways (Figure 3).

Data mining is one of the steps in the iterative process of knowledge discovery and it is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data and it consist of applying data analysis and discovery (learning) algorithms that produce a particular enumeration of patterns (or models) over the data (Cios & Moore 2002). Data mining combines techniques such as statistical analysis, visualization and induction to explore large amounts of data and discover relationships and patterns that shed light on periodontitis problems. Data mining would also help analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed and thus predict aspects of the disease, that may allow researches and clinicians to make proactive, knowledge-driven decisions. The analytical techniques used in data mining are often well-known mathematical algorithms and techniques like artificial neural networks, decision trees, rule induction or nearest neighbor.

There are problems associated with the process of data mining that often suggest a certain class of data mining technique or method to be an appropriate solution.

- **Clustering** is often one of the first steps in medical areas. It identifies groups of related records that can be used as a starting point for exploring further relationships. Clustering techniques are frequently used to discover structure or similarities in data (Gordon 1987, Michalski & Stepp 1983). In the health care profession, this type of problem is especially interesting to researchers and health care insurers or providers trying to discover information about a drug, a treatment or a disease.

The classification of periodontitis is still a matter of debate due to its complexity and the subjective criteria used by clinicians; the heterogeneity in diagnostics prevent an objective and general evaluation of the disease. For this reason it would be interesting to be able to define objectively the states or types of this disease and in such a way to unify the criteria of classification. This would be possible to realize across the clustering of clinical data.

- **Feature subjects selection (FSS)** reduces the number of statistical features in high-dimensional classification problems (Kohavi & John 1997). In the study of this complex disease that involves multiple variables, FSS could help clinicians by finding subsets of features that are most relevant. Models constructed in the subspace of the selected features tend to generalize new data better than classifiers trained on the entire feature set.
- **Classification.** This problem involves the need to find relationships that can partition the data into disjoint groups. Classification, perhaps the most commonly applied data mining technique, employs a set of labeled examples to develop a model that can classify the unlabeled population (Duda & Hart 1973, Fisher 1936). Example-based data mining methods such as nearest neighbor classification, regression algorithms and case-based reasoning are also examples of solutions to data classification problems. For periodontitis health care professionals, this type of data mining technique would be important in diagnostic decision making since by these probabilistic methods, the clinicians might obtain the probabilities that a given patient is suffering from a certain state of periodontitis.

- **Association** tries to find all of the rules (or at least a critical subset of rules) for which a particular data attribute is either a consequence or an antecedent of interest to health care professionals who are looking for relationships between (i) diseases and life-styles and demographic variables; (ii) between survival rates and treatments. An example of such a technique is causal network. Often association-type data mining techniques are employed to help strengthen arguments concerning whether or not to include or eliminate candidate rules from a knowledge model.

In the future more studies can be generated since the variables studied in periodontitis are mostly biological parameters interconnected. In this new discipline, biological pathways can be constructed through the use of computational algorithms like bayesian networks to generate meaningful data from gene, protein-protein interaction studies, and other experiments. This will enable us to understand how parameters are interconnected, to identify therapeutically relevant targets, and define and diagnose disease on a molecular basis. Data architectures underlie the accurate diagnosis and early intervention of disease, based on genotype, gene expression signatures and protein transcription. This knowledge will help deliver personalized medicine and lower the cost of drug development.

Data Visualization. Data visualization software is one of the tools for data mining interpretation. It enables you to visually interpret complex patterns in multidimensional data. By viewing data summarized in multiple graphical forms and dimensions, you can uncover trends and spot outliers intuitively and immediately. Visualization can be a great aid in the early stages of pattern discovery, to help “sanitary checking” in data preparation, verify correctness of preprocessing steps, clean up undesirable artifacts, and choose relevant samples. Visualization also enables an initial exploration to spot explicit patterns, select potentially useful features, try different normalization schemes, and in the future will suggest choices of classifiers, clustering algorithms, or trend models that are included in metalearning developing.

In the data mining process, visualization tools help to explore data before modeling and to verify the results of other data mining techniques. Visualization tools are particularly useful for detecting patterns found in only small areas of the overall data. Understandably it is difficult for a single visualization method to satisfy all these needs. Ideally, data visualization should allow large flexibility in the choice of perspectives, ranging from basic statistical graphics such as histograms

and scatter plots, to more sophisticated views of high dimensional spaces, multimedia signals, maps, and auxiliary data structures like graphs and trees. Many innovative data views have been produced in visualization research. However, most visualization systems are implemented as a toolkit, i.e., a loose collection of a library of plotting modules. Better uses of these can be made if there is a connecting architecture, where data relationships can be easily tracked between modules.

Good data mining techniques are faced with a series of extremely difficult challenges. Some of these are: high dimensionality (very large number of attributes to compare and explore for relationships); missing, incomplete and noisy (or inaccurate) data; overfitting (this is a particular problem for classification); extremely complex relationships between variables that simplistic techniques cannot detect; integration between data sources and data types; volatility of some of the data and knowledge; assessing the statistical significance of the patterns detected; the impact on the results that data preprocessing has; the visualization issue that plagues result interpretation and the lack of human ability to understand some of the complex patterns that a computer can detect; privacy issues (especially relevant to medical information); and the meaningfulness of the patterns detected.

4. Relevance to the study of chronic inflammatory diseases

The management and control of chronic inflammatory diseases are a major challenge for the health systems of all the European countries. Since these diseases are multifactorial and polygenic it is likely that before the etiology is known, a classification of these diseases based on molecular biology, genetic information, phenotype information and modern imaging systems, are bound to provide the clinicians with new tools to identify the different subgroups with different prognosis. Based on these classifications different strategies can be developed for additional diagnosis, therapy and patient management. In this new scenario, old and new treatment strategies can be tested in randomized genomic-based clinical trials using the newly proposed classifications (diagnostic systems) and patient subgroups.

Data collected using new approaches in the field of BMI will permit to build large datasets that can be analyzed. The development and implication of new data warehouse tools that become available for clinicians and researchers, will allow them to perform newly developed techniques in

data mining and visualization. New tools in these fields will help in metalearning. This opens up new steps towards understanding the (relative) roles of various proposed etiologic factors. The current case study, i.e. periodontitis, will serve as a model for other chronic inflammatory diseases. In this pilot project, periodontitis was selected because it provides an example of a chronic, complex disease where genetic causes are surely involved. Compared to other complex diseases, such as cancer or diabetes, where most likely a much larger number of genes and a plethora of environmental factors might be involved, periodontitis favors a reduced research approach. Focus on a limited number of etiological factors is likely to lead to the discovery of new clinico-genetic associations and scientific hypotheses, which can be tested. Similarly, new advances in the pathophysiology of periodontitis may provide experience and directions for many other multifactorial/complex diseases.

5. Conclusions

In this paper we have analyzed the area of complex diseases and its implications for future informatics-based approaches to research, patient care and management. Periodontitis is an ideal target for proving the feasibility of such approaches, given the specific clinical and genetic information that can be analyzed. Considering BI or MI, until recently a myriad of methods was applied; the current study of periodontitis within the Network of Excellence “INFOBIOMED” is an example of integrative research approach, where the new field of BMI will contribute to new insights in chronic inflammatory and infectious diseases.

To design new approaches in the study of complex diseases such as periodontitis, to develop better classification schemes and to make strong breakthroughs in diagnosis, prevention, prognosis, treatment issues and even pharmaceutical approaches, we need to bring together periodontology, MI and BI. This will be successful if all data are generated, stored and handled uniformly, according to European standards. This concerns not only semantics, but more importantly also syntax of data. Protocols for data transmission via HL7 or EDIFACT need to be generated for biomedical exchange, taking privacy issues into consideration. The INFOBIOMED consortium is currently developing a number of tools for these purposes, using periodontitis as model for the study of complex diseases.

References

- Al-Zahrani, M. S., Bissada, N. F. & Borawski, E. A. (2005) Diet and periodontitis. *J Int Acad Periodontol* **7**, 21-26.
- Armitage, G. C. (1999) Development of a classification system for periodontal diseases and conditions. *Ann Periodontol* **4**, 1-6.
- Beck, J. D., Pankow, J., Tyroler, H. A. & Offenbacher, S. (1999) Dental infections and atherosclerosis. *Am Heart J* **138**, S528-533.
- Berglundh, T., Donati, M., Hahn-Zoric, M., Hanson, L. A. & Padyukov, L. (2003) Association of the -1087 IL 10 gene polymorphism with severe chronic periodontitis in Swedish Caucasians. *J Clin Periodontol* **30**, 249-254.
- Bergström, J. (1989) Cigarette smoking as risk factor in chronic periodontal disease. *Com Dent Oral Epidemiol* **17**, 245-247.
- Bostrom, L., Bergstrom, J., Dahlen, G. & Linder, L. E. (2001) Smoking and subgingival microflora in periodontal disease. *J Clin Periodontol* **28**, 212-219.
- Bouclin, R., Landry, R. G. & Noreau, G. (1997) The effects of smoking on periodontal structures: a literature review. *J Can Dent Assoc* **63**, 356, 360-353.
- Boughman, J. A., Astemborski, J. A. & Suzuki, J. B. (1992) Phenotypic assessment of early onset periodontitis in sibships. *J Clin Periodontol* **19**, 233-239.
- Breivik, T., Thrane, P. S., Murison, R. & Gjermo, P. (1996) Emotional stress effects on immunity, gingivitis and periodontitis. *Eur J Oral Sci* **104**, 327-334.
- Brookes, A. J. (2001) Rethinking genetic strategies to study complex diseases. *Trends Mol Med* **7**, 512-516.
- Brown, L. J., Oliver, R. C. & Loe, H. (1990) Evaluating periodontal status of US employed adults. *J Am Dent Assoc* **121**, 226-232.
- Calsina, G., Ramon, J. M. & Echeverria, J. J. (2002) Effects of smoking on periodontal tissues. *J Clin Periodontol* **29**, 771-776.
- Chapple, I. L. (1997) Reactive oxygen species and antioxidants in inflammatory diseases. *J Clin Periodontol* **24**, 287-296.
- Chen, C. (2001) Periodontitis as a biofilm infection. *J Calif Dent Assoc* **29**, 362-369.
- Cios, K. J. & Moore, G. W. (2002) Uniqueness of medical data mining. *Artif Intell Med* **26**, 1-24.
- Clayton, D. & McKeigue, P. M. (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* **358**, 1356-1360.

- Colhoun, H. M., McKeigue, P. M. & Davey Smith, G. (2003) Problems of reporting genetic associations with complex outcomes. *Lancet* **361**, 865-872.
- de Brito Junior, R. B., Scarel-Caminaga, R. M., Trevilatto, P. C., de Souza, A. P. & Barros, S. P. (2004) Polymorphisms in the vitamin D receptor gene are associated with periodontal disease. *J Periodontol* **75**, 1090-1095.
- De Moor, G. J. E., Claerhout, B. & De Meyer, F. (2003) Privacy enhancing techniques: the key to secure communication and management of clinical and genomic data. *Methods Inf Med* **42**, 148-153.
- Duda, R. & Hart, P. Pattern classification and scene analysis. In, eds. New York: John Wiley & Sons; 1973:
- Fisher, R. (1936) The use of multiple measurements. *Annals of Eugenics* **7**, 179-188.
- Gjeramo, P. (1998) Epidemiology of periodontal diseases in Europe. *J Parodontol d'Implantol Orale* **17**, 111-121.
- Gonzales, J. R., Michel, J., Diete, A., Herrmann, J. M., Bodeker, R. H. & Meyle, J. (2002) Analysis of genetic polymorphisms at the interleukin-10 loci in aggressive and chronic periodontitis. *J Clin Periodontol* **29**, 816-822.
- Gordon, A. D. (1987) A review of hierarchical classification. *Journal of the Royal Statistical Society Series A*, **150**, 119-137.
- Graswinckel, J. E., van der Velden, U., van Winkelhoff, A. J., Hoek, F. J. & Loos, B. G. (2004) Plasma antibody levels in periodontitis patients and controls. *J Clin Periodontol* **31**, 562-568.
- Griffen, A. L., Becker, M. R., Lyons, S. R., Moeschberger, M. L. & Leys, E. J. (1998) Prevalence of Porphyromonas gingivalis and periodontal health status. *J Clin Microbiol* **36**, 3239-3242.
- Grossi, S. G., Skrepcinski, F. B., DeCaro, T., Zambon, J. J., Cummins, D. & Genco, R. J. (1996) Response to periodontal therapy in diabetics and smokers. *J Periodontol* **67**, 1094-1102.
- Haber, J., Wattles, J., Crowlery, M., Mandell, R., Joshipura, K. & Kent, R. L. (1993) Evidence for cigarette smoking as a major risk factor for periodontitis. *J Periodontol* **64**, 16-23.
- Hart, T. C., Hart, P. S., Bowden, D. W., Michalec, M. D., Callison, S. A., Walker, S. J., Zhang, Y. & Firatli, E. (1999) Mutations of the cathepsin C gene are responsible for Papillon-Lefevre syndrome. *J Med Genet* **36**, 881-887.
- Hart, T. C., Hart, P. S., Michalec, M. D., Zhang, Y., Marazita, M. L., Cooper, M., Yassin, O. M., Nusier, M. & Walker, S. (2000) Localisation of a gene for prepubertal periodontitis to chromosome 11q14 and identification of a cathepsin C gene mutation. *J Med Genet* **37**, 95-101.

- Hennig, B. J., Parkhill, J. M., Chapple, I. L., Heasman, P. A. & Taylor, J. J. (1999) Association of a vitamin D receptor gene polymorphism with localized early-onset periodontal diseases. *J Periodontol* **70**, 1032-1038.
- Hewitt, C., McCormick, D., Linden, G., Turk, D., Stern, I., Wallace, I., Southern, L., Zhang, L., Howard, R., Bullon, P., Wong, M., Widmer, R., Gaffar, K. A., Awawdeh, L., Briggs, J., Yaghmai, R., Jabs, E. W., Hoeger, P., Bleck, O., Rudiger, S. G., Petersilka, G., Battino, M., Brett, P., Hattab, F., Al-Hamed, M., Sloan, P., Toomes, C., Dixon, M., James, J., Read, A. P. & Thakker, N. (2004) The role of cathepsin C in Papillon-Lefevre syndrome, prepubertal periodontitis, and aggressive periodontitis. *Hum Mutat* **23**, 222-228.
- Ioannidis, J. P. (2003) Genetic associations: false or true? *Trends Mol Med* **9**, 135-138.
- Ioannidis, J. P., Trikalinos, T. A., Ntzani, E. E. & Contopoulos-Ioannidis, D. G. (2003) Genetic associations in large versus small studies: an empirical assessment. *Lancet* **361**, 567-571.
- Kinane, D. F. & Hart, T. C. (2003) Genes and gene polymorphisms associated with periodontal disease. *Crit Rev Oral Biol Med* **14**, 430-449.
- Kohavi, R. & John, G. (1997) Wrappers for feature subset selection. *Artificial Intelligence* **97**, 273-324.
- Kornman, K. S., Crane, A., Wang, H. Y., di Giovine, F. S., Newman, M. G., Pirk, F. W., Wilson, T. G., Jr., Higginbottom, F. L. & Duff, G. W. (1997) The interleukin-1 genotype as a severity factor in adult periodontal disease. *J Clin Periodontol* **24**, 72-77.
- Kuipers, J. G., Zeidler, H. & Kohler, L. (2003) How does Chlamydia cause arthritis? *Rheum Dis Clin North Am* **29**, 613-629.
- Laine, M. L., Farre, M. A., Gonzalez, G., van Dijk, L. J., Ham, A. J., Winkel, E. G., Crusius, J. B., Vandenbroucke, J. P., van Winkelhoff, A. J. & Pena, A. S. (2001) Polymorphisms of the interleukin-1 gene family, oral microbial pathogens, and smoking in adult periodontitis. *J Dent Res* **80**, 1695-1699.
- Liuba, P. (2004) Arterial endothelial injury due to infection in childhood: ticking bomb or innocent bystander? *Acta Paediatr Suppl* **93**, 55-62.
- Loos, B. G., Hamming, H. & Van der Velden, U. (1998) Stress and periodontitis: a literature review. *Journal de Parodontologie & d' Implantologie Orale* **17**, 205-217.
- Loos, B. G., Leppers-Van de Straat, F. G., Van de Winkel, J. G. & Van der Velden, U. (2003) Fcgamma receptor polymorphisms in relation to periodontitis. *J Clin Periodontol* **30**, 595-602.
- Loos, B. G., Roos, M. T., Schellekens, P. T., van der Velden, U. & Miedema, F. (2004) Lymphocyte numbers and function in relation to periodontitis and smoking. *J Periodontol* **75**, 557-564.
- Loos, B. G. (2005) Systemic markers of inflammation in periodontitis. *J Periodontol* **76**, in press.

- Loos, B. G., John, R. P. & Laine, M. L. (2005) Identification of genetic risk factors for periodontitis and possible mechanisms of action. *J Clin Periodontol* **32 Suppl 6**, 159-179.
- MacFarlane, G. D., Herzberg, M. C., Wolff, L. F. & Hardie, N. A. (1992) Refractory periodontitis associated with abnormal polymorphonuclear leukocyte phagocytosis and cigarette smoking. *J Periodontol* **63**, 908-913.
- Martin-Sanchez, F., Iakovidis, I., Norager, S., Maojo, V., de Groen, P., Van der Lei, J., Jones, T., Abraham-Fuchs, K., Apweiler, R., Babic, A., Baud, R., Breton, V., Cinquin, P., Doupi, P., Dugas, M., Eils, R., Engelbrecht, R., Ghazal, P., Jehenson, P., Kulikowski, C., Lampe, K., De Moor, G., Orphanoudakis, S., Rossing, N., Sarachan, B., Sousa, A., Spekowius, G., Thireos, G., Zahlmann, G., Zvarova, J., Hermosilla, I. & Vicente, F. J. (2004) Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *J Biomed Inform* **37**, 30-42.
- Meisel, P., Carlsson, L. E., Sawaf, H., Fanghaenel, J., Greinacher, A. & Kocher, T. (2001) Polymorphisms of Fc γ -receptors RIIa, RIIIa, and RIIIb in patients with adult periodontal diseases. *Genes Immun* **2**, 258-262.
- Michalowicz, B. S., Aeppli, D., Virag, J. G., Klump, D. G., Hinrichs, J. E., Segal, N. L., Bouchard, T. J. & Philstrom, B. L. (1991) Periodontal findings in adult twins. *J Periodontol* **62**, 293-299.
- Michalowicz, B. S., Diehl, S. R., Gunsolley, J. C., Sparks, B. S., Brooks, C. N., Koertge, T. E., Califano, J. V., Burmeister, J. A. & Schenkein, H. A. (2000) Evidence of a substantial genetic basis for risk of adult periodontitis. *J Periodontol* **71**, 1699-1707.
- Michalski, R. S. & Stepp, R. E. (1983) Automated construction of classifications: Conceptual clustering versus numerical taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5**, 219-243.
- Mukamal, K. J., Kronmal, R. A., Tracy, R. P., Cushman, M. & Siscovick, D. S. (2004) Traditional and novel risk factors in older adults: cardiovascular risk assessment late in life. *Am J Geriatr Cardiol* **13**, 69-80.
- Page, R. C., Offenbacher, S., Schroeder, H. E., Seymour, G. J. & Kornman, K. S. (1997) Advances in the pathogenesis of periodontitis: summary of developments, clinical implications and future directions. *Periodontol 2000* **14**, 216-248.
- Palmer, R. M., Wilson, R. F., Hasan, A. S. & Scott, D. A. (2005) Mechanisms of action of environmental factors--tobacco smoking. *J Clin Periodontol* **32 Suppl 6**, 180-195.
- Papapanou, P. N. & Lindhe, J. Epidemiology of periodontal diseases. In: Lindhe, J., Karring, T. & Lang, N. P., eds. *Clinical Periodontology and Implant Dentistry*, 4th Ed. Oxford: Blackwell Munksgaard; 2003:50-80.
- Park, A. (2002) Beyond cholesterol. Inflammation is emerging as a major risk factor--and not just in heart disease. *Time* **160**, 74-75.

- Picolos, D. K., Lerche-Sehm, J., Abron, A., Fine, J. B. & Papapanou, P. N. (2005) Infection patterns in chronic and aggressive periodontitis. *J Clin Periodontol* **32**, 1055-1061.
- Sanz, F., Diaz, C., Martin-Sanchez, F. & Maojo, V. (2004) Structuring European biomedical informatics to support individualized healthcare: current issues and future trends. *Medinfo* **11**, 803-807.
- Scarel-Caminaga, R. M., Trevilatto, P. C., Souza, A. P., Brito, R. B., Camargo, L. E. & Line, S. R. (2004) Interleukin 10 gene promoter polymorphisms are associated with chronic periodontitis. *J Clin Periodontol* **31**, 443-448.
- Slots, J. & Ting, M. (2002) Systemic antibiotics in the treatment of periodontal disease. *Periodontol* **2000** **28**, 106-176.
- Socransky, S. S., Haffajee, A. D., Cugini, M. A., Smith, C. & Kent, R. L., Jr. (1998) Microbial complexes in subgingival plaque. *J Clin Periodontol* **25**, 134-144.
- Socransky, S. S., Smith, C. & Haffajee, A. D. (2002) Subgingival microbial profiles in refractory periodontal disease. *J Clin Periodontol* **29**, 260-268.
- Staudte, H., Sigusch, B. W. & Glockmann, E. (2005) Grapefruit consumption improves vitamin C status in periodontitis patients. *Br Dent J* **199**, 213-217, discussion 210.
- Sujansky, W. (2001) Heterogeneous database integration in biomedicine. *J Biomed Inform* **34**, 285-298.
- Sun, J. L., Meng, H. X., Cao, C. F., Tachi, Y., Shinohara, M., Ueda, M., Imai, H. & Ohura, K. (2002) Relationship between vitamin D receptor gene polymorphism and periodontitis. *J Periodontal Res* **37**, 263-267.
- Tabor, H. K., Risch, N. J. & Myers, R. M. (2002) Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* **3**, 391-397.
- Toomes, C., James, J., Wood, A. J., Wu, C. L., McCormick, D., Lench, N., Hewitt, C., Moynihan, L., Roberts, E., Woods, C. G., Markham, A., Wong, M., Widmer, R., Ghaffar, K. A., Pemberton, M., Hussein, I. R., Temtamy, S. A., Davies, R., Read, A. P., Sloan, P., Dixon, M. J. & Thakker, N. S. (1999) Loss-of-function mutations in the cathepsin C gene result in periodontal disease and palmoplantar keratosis. *Nat Genet* **23**, 421-424.
- Van der Velden, U., Abbas, F., Armand, S., de Graaff, J., Timmerman, M. F., van der Weijden, G. A., van Winkelhoff, A. J. & Winkel, E. G. (1993) The effect of sibling relationship on the periodontal condition. *J Clin Periodontol* **20**, 683-690.
- Van der Velden, U. (2000) Diagnosis of periodontitis. *J Clin Periodontol* **27**, 960-961.
- Van der Velden, U. (2005) Purpose and problems of periodontal disease classification. *Periodontol* **2000** **39**, 13-21.

- Van Winkelhoff, A. J., Bosch-Tijhof, C. J., Winkel, E. G. & van der Reijden, W. A. (2001) Smoking affects the subgingival microflora in periodontitis. *J Periodontol* **72**, 666-671.
- van Winkelhoff, A. J., Loos, B. G., van der Reijden, W. A. & van der Velden, U. (2002) *Porphyromonas gingivalis*, *Bacteroides forsythus* and other putative periodontal pathogens in subjects with and without periodontal destruction. *J Clin Periodontol* **29**, 1023-1028.
- Xu, L., Loos, B. G., Craandijk, J., Ritsema, E., Huffels, R. A. M. & Van der Velden, U. (2002) Teeth with periodontal bone loss, cigarette smoking and plasma cotinine levels. *J Intern Acad Periodontol* **4**, 39-43.
- Yamamoto, K., Kobayashi, T., Grossi, S., Ho, A. W., Genco, R. J., Yoshie, H. & De Nardin, E. (2004) Association of Fcγ receptor IIa genotype with chronic periodontitis in Caucasians. *J Periodontol* **75**, 517-522.

Table 1. Summary of subgingival bacterial species associated with periodontitis and their position in proposed microbial complexes (Socransky et al. 1998).

Complex	Species
Red	<i>Porphyromonas gingivalis</i> <i>Treponema denticola</i> <i>Bacteroides forsythus</i>
Orange	<i>Streptococcus constellatus</i> <i>Eubacterium nodatum</i> <i>Fusobacterium nucleatum ss vincentii</i> <i>Campylobacter rectus</i> <i>Peptostreptococcus micros</i> <i>Prevotella nigrescens</i> <i>Fusobacterium nucleatum ss polymorphum</i> <i>Campylobacter showae</i> <i>Fusobacterium periodonticum</i> <i>Fusobacterium nucleatum ss nucleatum</i> <i>Campylobacter gracilis</i> <i>Prevotella intermedia</i>
Green	<i>Actinobacillus actinomycetemcomitans serotype a</i> <i>Capnocytophaga ochracea</i> <i>Campylobacter concisus</i> <i>Capnocytophaga gingivalis</i> <i>Capnocytophaga sputigena</i> <i>Eikenella corrodens</i> <i>Capnocytophaga species</i>
Purple	<i>Actinomyces odontolyticus</i> <i>Veilonela parvula</i>
Yellow	<i>Streptococcus sanguis</i> <i>Streptococcus oralis</i> <i>Streptococcus intermedius</i> <i>Streptococcus gordonii</i> <i>Streptococcus species</i> <i>Streptococcus mitis</i>
Blue	<i>Actinomyces species</i>
Miscellaneous	<i>Actinobacillus actinomycetemcomitans serotype b</i> <i>Bacteroides fragilis</i> <i>Campylobacter sputorum ss bubulus</i> <i>Bacteroides ureolyticus</i> <i>Campylobacter sputorum ss sputorum</i> <i>Campylobacter curvus</i> <i>Selenomonas noxia</i> <i>Porphyromonas endodontalis</i> <i>Wollinella succinogenes</i>

Table 2. Summary of candidate genes, and the corresponding encoded proteins, for which gene polymorphisms have been investigated as putative risk factors for periodontitis (Loos et al. 2005).

Gene	Coded protein
<i>ACE</i>	Angiotensin-converting enzyme
<i>CARD15 (NOD2)</i>	Caspase recruitment domain-15
<i>CCR5</i>	Chemokine receptor-5
<i>CD14</i>	CD-14
<i>ER2</i>	Estrogen receptor-2
<i>ET1</i>	Endothelin-1
<i>FBR</i>	Fibrinogen
<i>FcγRIIa</i>	Fc γ receptor IIa
<i>FcγRIIb</i>	Fc γ receptor IIb
<i>FcγRIIIa</i>	Fc γ receptor IIIa
<i>FcγRIIIb</i>	Fc γ receptor IIIb
<i>FPR1</i>	N-formylpeptide receptor-1
<i>IFNGR1</i>	Interferon γ receptor-1
<i>IL1A</i>	Interleukin-1α
<i>IL1B</i>	Interleukin-1β
<i>IL1RN</i>	Interleukin-1 receptor antagonist
<i>IL2</i>	Interleukin-2
<i>IL4</i>	Interleukin-4
<i>IL6</i>	Interleukin-6
<i>IL10</i>	Interleukin-10
<i>LTA</i>	Lymphotoxin-α
<i>MMP1</i>	Matrix metalloproteinase-1
<i>MMP3</i>	Matrix metalloproteinase-3
<i>MMP9</i>	Matrix metalloproteinase-9
<i>MPO</i>	Myeloperoxidase
<i>NAT2</i>	N-acetyltransferase-2
<i>PAI1</i>	Plasminogen-activator-inhibitor-1
<i>RAGE</i>	Receptor for advanced glycation end products
<i>TGFB</i>	Transforming growth factor-β
<i>TIMP2</i>	Tissue inhibitor of matrix metalloproteinase
<i>TLR2</i>	Toll-like receptor-2
<i>TLR4</i>	Toll-like receptor-2
<i>TNFA</i>	Tumor necrosis factor-α
<i>TNFR2</i>	Tumor necrosis factor receptor-2
<i>VDR</i>	Vitamin D receptor

Table 3. This table shows the multiple and heterogeneous data that will be collected for the periodontitis data warehouse (PDW).

Group of data	Example	Collection place and method	External data sources
Disease Phenotype	Teeth loss Severity	Dental clinic/hospital, X-rays, interviews, clinical examination Analytical labs	Clinical Patient record (CPR) PACS, Laboratory information system
Intermediate Phenotypes	Age		
	Gender		
	Weight		
	Height		
	Body Mass Index		
	Blood pressure		
	Biochemical parameters (triglycerides)		
Glucose			
Genotype	SNPs, Haplotypes	Genetics lab (sequencing, PCR)	dbSNP, OMIM HGNC, Entrez Gene, GO
	Mutations		
	Genes		
Infectious agent	Endogenous bacteria	Microbiology lab (culture, antibiogram)	NCBI Taxonomy / Bacterial Nomenclature DSMZ
	Exogenous Bacteria		
Environmental factors, life habits and history of drug consumption.	Smoking	Questionnaires	-
	Diet (vitamins)		
	Stress		
	Oral Hygiene		
	Drugs	Clinical interview or CPR	Drugs DB
	Demographic data	Statistics	Epidemiology databases

Legends to figures

Figure 1.

Example of a dental radiograph of tooth 31 and surrounding hard tissue structures showing important landmarks: the cementoenamel junction (CEJ), the apex and the periodontal bone level (BL) on the mesial site of the tooth. Periodontal bone levels in the healthy situation are located 1-2 mm apical from the CEJ; in this example there is about 55% periodontal bone lost on the mesial surface.

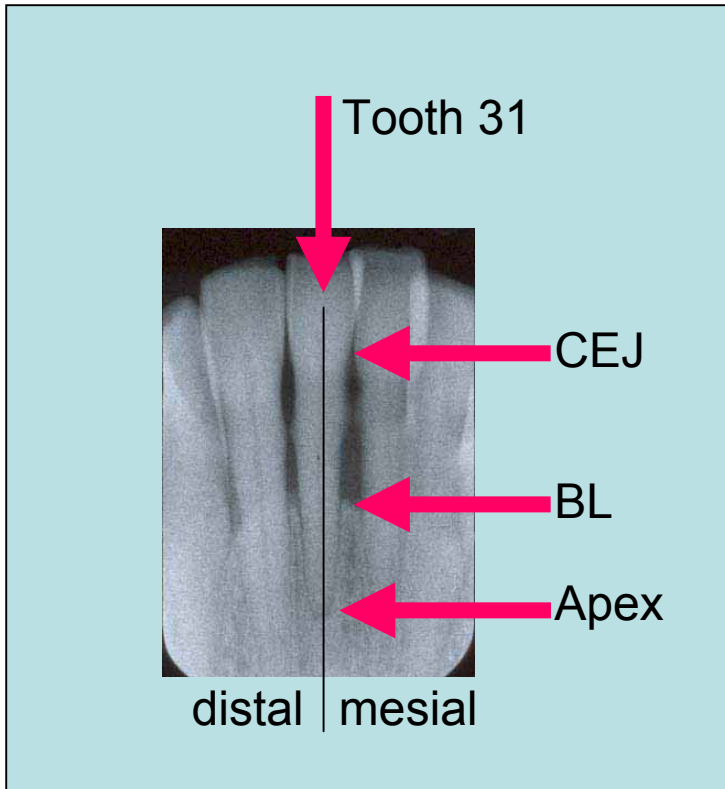
Figure 2.

Technological needs for BMI in periodontitis

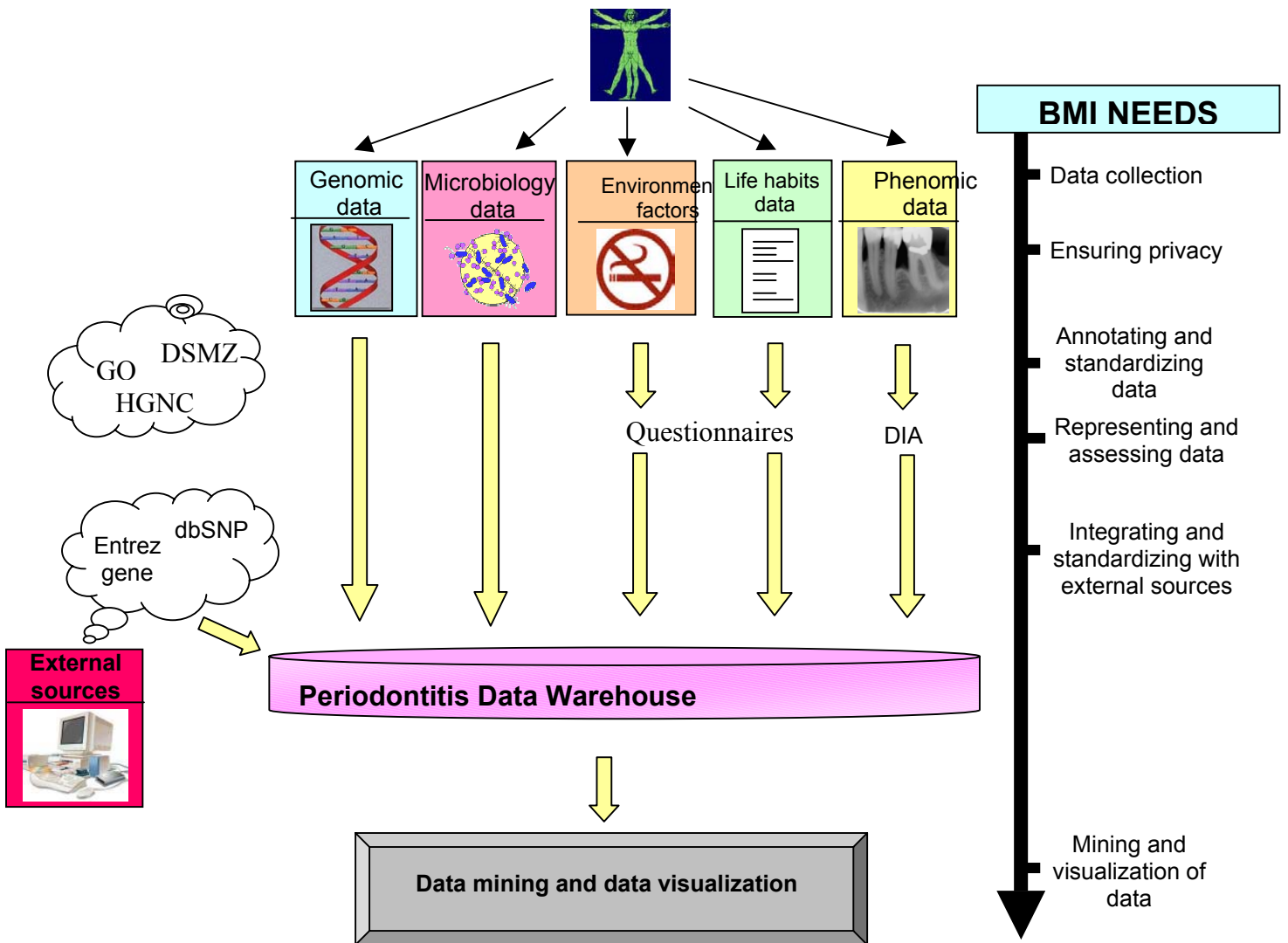
Figure 3.

The data mining process.

(Figure 1)



(Figure 2)



(Figure 3)

