Design of a Global Medical Database which is Searchable by Human Diagnostic Patterns

Wolfgang Orthuber[*,1], Gunar Fiedler[2], Michael Kattan[3], Thorsten Sommer[1] and Helge Fischer-Brandies[1]

[1]*Department of Orthodontics at University of Kiel, Germany*

[2]*Institute of Informatics at University of Kiel, Germany*

[3]*Department of Quantitative Health Sciences, Cleveland Clinic Foundation, Cleveland, OH, USA*

Abstract: We describe a global medical database which is designed for efficient evaluation. It allows language independent search for human diagnostic parameters. Core of the database is a fully automated electronic archive and distribution server for medical histories of real but anonymous patients which contain patterns of diagnosis, chosen treatment, and outcome. Every pattern is represented by a feature vector which is usually a sequence of numbers, and labeled by an unambiguous "pattern name" which identifies its meaning. Similarity search is always done only over patterns with the same pattern name, because these are directly comparable. Similarities of patterns are mapped to spatial similarities (small distances) of their feature vectors using an appropriate metric. This makes them searchable. Pattern names can be "owned" like today domain names. This facilitates unbureaucratic definition of patterns e.g. by manufacturers of diagnostic devices. Application: If there is a new patient with certain diagnostic patterns, it is possible to combine a part or all of them and to search in the database for completed histories of patients with similar patterns to find the best treatment. Confinement of the result by conventional language based search terms is possible, and immediate individual statistics or regression analyses can quantify probabilities of success in case of different treatment choices. Conclusions: Efficient searching with diagnostic patterns is technically feasible. Labeled feature vectors induce a systematic and expandable approach. The database also allows immediate calculation of individual up to date prediction models.

INTRODUCTION

Every conscientious doctor is aware of the boundary of the own local experience and appreciates possibilities to expand it. Advanced training can help, but even for specialists it would simply need too much time to cover without relevant simplifications the increasing complexity of all possible measurements, diagnoses and therapies. So additional decision aids are necessary, concretely for given diagnostic results the doctor needs to know possible therapies, their long term consequences and probabilities of success in case of this or that therapy. For these purposes prediction tools have been developed which use the experience from thousands of patients. These are of particular importance in case of clinical decisions with serious consequences. For example in case of cancer it is necessary to choose a therapy which avoids stoppable progression as well as unnecessary trauma. Alone for prostate cancer numerous tools have been developed to make adequate predictions, for example neural networks [1], probability tables [2] and nomograms [3-8], which are also available for sarcoma [9], melanoma [10], gastric carcinoma [11], renal cell carcinoma [12], cancer of breast [13], bladder [14], lung [15] and pancreas [16]. Here we can mention only a few examples, today there are so many models, that for some situations selection becomes difficult.

*Address correspondence to this author at the Department of Orthodontics at University of Kiel, Germany; E-mail: g51@orthuber.com

All these models are derived from collections of patient histories. Meanwhile the web allows the more efficient direct way: To store all these and further documentations in standardized form in a global database which is searchable by numerical patterns. At this every pattern is represented by a feature vector which is usually a sequence of numbers, and labeled by an unambiguous "pattern name" which identifies its meaning. From such a database one could update and refine continuously given models and develop new models. It could be also used directly by the doctor to find for a given patient clinical records of patients with similar diagnostic patterns. These could contain the *completed* story *after* some chosen treatment and detailed specific and valuable advices of the colleagues, an immediate individual statistics over these "similar patients" or even a complete regression analysis using all records with patterns of this kind can be calculated which allows to estimate probabilities of success in case of different treatment choices. So valid decision help and implicitly a free individual advanced training would be always possible if somewhere similar cases have been documented and uploaded. The more specific the diagnostic measurements are, the better. Typical searchable diagnostic patterns may be:

- Results of blood tests, e.g. concentrations of tumor associated antigens like PSA, f-PSA etc.
- Anamnestic data like age, gender, BMI; also body height can be relevant, e.g. in orthopedics.

1874-4311/08

- Other relevant diagnostic measurements, appropriately preprocessed, e.g. MRI scans after feature extraction [17].

There would be additional motivation for the manufacturer to provide diagnostic means which produce highly selective and reproducible results. This shows that there is much commercial potential, too.

Concise Questions and Answers

Due to the variability of diagnostic measures resp. patterns it is a comprehensive task to make them searchable and requires a clear reply to the following questions:

- How can the doctor provide a diagnostic pattern for which similar patterns should be found on the database?
- How can the database recognize the kind of diagnostic pattern provided by the doctor and so isolate the set of comparable patterns on the database?
- How can the database quantify the similarity between the provided pattern and the comparable patterns with attached articles in the archive to calculate their rank in the search result?

It turns out that there are satisfying answers to these questions in all cases of comparable patterns. They are abbreviated:

In case of concise patterns the doctor can enter the pattern directly by keyboard, e.g. as sequence of numbers, together with an unique "pattern name", which specifies the kind of the pattern. In other cases the doctor can upload the pattern as file from diagnostic software which is designed for handling of this pattern kind. If appropriate, this software may be connected with some digitizing device or it uses data from a laboratory.

The database recognizes the kind of the pattern by the pattern name. Then it searches within the archive for publications whose associated patterns have the same pattern name. Those with most similar numerical representation will have highest rank in the search result.

Quantification of similarity depends on the kind of the pattern, which is known together with the pattern name. The numerical representation of the pattern is designed in a way that the database can calculate their similarity by direct comparison using an efficient algorithm.

ORGANIZATION OF THE DATABASE IN DETAIL

The database supports the universal pattern search concept [18] which could be also applied to the total web. It is arbitrarily expandable, every kind of pattern is represented by a feature vector which is a sequence of numbers, and labeled by an unambiguous name, the pattern name: Because arbitrarily many such names are conceivable, it is also possible to define arbitrarily many different kinds of patterns. Only patterns with the same pattern name are comparable using a metric with a short nonnegative distance function, e.g. Manhattan distance, Euclidean distance etc. [19]. For every pattern name the feature vector and the associated distance function can be individually defined. This means that any kind of pattern has its individual similarity criterion. Two patterns with the same pattern name are the more similar, the smaller the distance between their feature vectors is. Identical patterns have zero distance.

The subsequently suggested organizational details should represent an efficient possibility for realization. Variants are conceivable. Important is that responsibility and necessary work are clearly shared in a way that it is attractive to participate.

Names and Conventions

First of all it is appropriate to explain some frequently used abbreviations:

Pattern

Some digital form of information. A searchable feature vector which is usually a sequence of numbers[1] represents it. The dimensionality of the feature vector (the length of the number sequence) is variable; it depends on the pattern definition. Similarities of the original data are mapped to spatial similarities of the feature vectors.

Pattern Files

Every pattern file represents a well-defined pattern. It contains a header with the pattern name which identifies the kind of a pattern, the date, a link to associated patient records or text, and some additional information, e.g. optionally a link to original data[2]. The feature vector which is the numerical representation of the pattern follows the header. We recommend a special filename ending for quick recognition and XML format as shown in [18].

Comparison of Patterns, Distance d

If a pattern is given for a search, it is necessary to quantify the similarity to other patterns on the database with the same pattern name. The result of such comparison is the deviation or distance $d \geq 0$, in which $d=0$ if the two compared patterns are identical, else $d>0$; d is the greater, the more they deviate. The smaller d, the higher is the rank of the associated URLs in the search result.

The multidimensional feature vectors which represent the patterns should be designed to be quickly comparable by the software of the database, using a short distance function for calculation of d.

Pattern Names and Pattern Domains

The *pattern name* is a string which can contain letters and numbers like Internet domain names, and points. It is the name which uniquely identifies the kind of a pattern; in the sense of the W3C it is an URN, a uniform resource name

[1] *All* findings can be converted into such a numerical form. To avoid misunderstandings definition and description of the ways of conversion is done in one place, in the pattern domain.

[2] Progress will continue, so also techniques for feature extraction will improve. If the numerical representation of a pattern contains features which are calculated from an original, e.g. from a picture, then also the original data should be uploaded and referred. So later advanced algorithms could be applied to calculate new additional searchable patterns.

[20]. To guarantee uniqueness also in case of expansion of the search method over the web we recommend the following convention:

Let *dn* denote the name of an Internet domain, in which patterns names are defined. Then these pattern names have the structure dn.*, where the ending * is a string without spaces. All definitions of patterns are done in the *pattern domain* which is a special subdirectory with name dn.pat (Fig. **2**). In our case dn is identical to the domain name of this database, if all definitions are done within it. Examples of pattern names may be:

"dn.blood-concentration.xyantigen",
"dn.blood-concentration.psapair",
"dn.ultrasonic.heart-results-1",
"dn.ekg1.avr",
"dn.vertebral-body-heights",
"dn.dna-seq.12",
"dn.features.fundus.oculi.1",
"dn.features.melanoma.4",
"dn.ICD.10",
"dn.evaluation.1" etc.

Pattern names make it possible to develop optimized structures and associated comparison algorithms to any kind of pattern independently of other kinds. Due to the variety of diagnostic methods and associated patterns it is necessary to share the work and to give motivation to participate. Therefore we orientate on the policy for Internet domain names which has been very successful. According to our suggestion the owner of the internet domain name dn owns also the pattern domain dn.pat and with this the privilege to define all pattern names of the form dn.*. The patterns with these names form a *pattern group*.

Motivation for Pattern Domain Owners

If a pattern group and domain should be useful and not ignored, its owner should:

- Provide useful definitions of all pattern components,
- If not trivial, describe efficient ways for their generation from original data,
- If necessary, give information to software for creation of patterns and/or donate or sell it,
- If necessary, give information about associated digitizing devices, he/she may also sell them.

Someone who invests much work in optimization of the own patterns can gain from this, because an efficient pattern is more frequently used. Some consequences:

- Communication in the own special field is more efficient.
- The own pattern domain "dn.pat" is more attractive
- The own software and/or digitizing devices which are necessary for generation of the dn.* patterns are more attractive.

So there are scientific *and* commercial reasons which make pattern domains attractive. The pattern domain owners play an important role; Fig. (**1**) illustrates the task sharing:

Motivation for Authors

Certainly there would be much motivation for a search request, if there are appropriate diagnostic digitizing devices and a good database.

But is there enough motivation for a doctor within today's framework to invest work and upload documentations resp. articles, to share own experiences with colleagues? Perhaps the feedback on this article will give first answers. At least the success of existing electronic archives indicates, that after some time of familiarization there can be also much motivation for doctors of medicine to become author in a worldwide read open archive. They can contribute a lot to science and progress by plain reality conform documentation. Health professionals who frequently upload, will become better known. Those who upload from the beginning

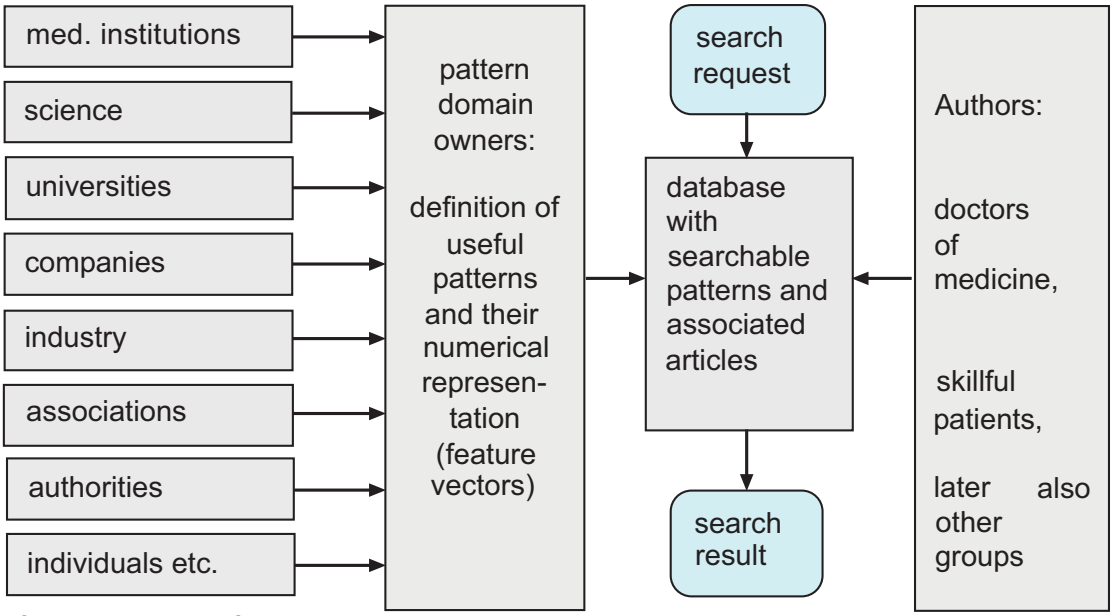

Fig. (1). Task sharing in pattern search.

will get particular attention, because initially there will not be a large number of authors, and all articles will be accessible chronologically.

There is another way for feeding the database which will become more and more important: Since some years there are increasing efforts [21] to collect all spread medical documentations of one patient in a standardized set of electronic medical records, which is accessible as a whole. After anonymization these records could be integrated into the database, if the patient explicitly whishes that. In this case he should also have the right to comment and to rate. Of course this would be an additional motivation for the doctors to achieve a good treatment result. If whished, the patient could be also contacted, e.g. for exchange of experiences in self help groups.

Appropriate Articles

Patient Histories

The archive will contain a directory with patient histories. These can be standardized electronic medical records, as mentioned above. Additionally all articles with honest documentations of medical treatments are welcome, also if the result of some treatment has been disappointing. If this is documented and uploaded, all colleagues could learn from it, repetition of ineffective treatment experiments could be avoided. We recommend liberal and user friendly possibilities for upload, because there is no danger of information overflow, if there are appropriate selective search options, e.g. individual restriction to certain groups of authors or qualities, according to the wishes of the user. Of course there must be minimal formal requirements which have been approved [22]. Additionally, there should be at least one searchable pattern in every treatment documentation, and the numerical representation of all patterns must be consistent. Under these circumstances also short articles are very welcome, e.g. records of patients who participate in clinical trials. So the results of clinical trials would be directly available and systematically searchable, further many misunderstandings could be avoided, because the today necessary detours with interpretations (this makes the reader of medical publications) of interpretations (this makes the author of medical publications, he interprets the clinical records) can be avoided. The database could make objective measurement results, and of course also intuitive quantitative ratings of patients broadly available and searchable for computers, without interpretation.

Documentations of Clinical Trials, Quality Descriptor

Additionally there will be a directory for conventional articles which describe clinical trials. The associated patient records should also be uploaded and referred. They will automatically get links backwards (Fig. **2**).

The author can designate his article as "xy percent documented", if in a prospective study the records of at least xy percent of initially participating patients are uploaded and evaluated. Then all referred patient records will get a mark which shows this quality. If one restricts a search request on records which belong to "100 percent documented" clinical trials, one has greater assurance that also patient records with unwanted outcomes are evaluated.

Values like xy could be integrated in a more general multidimensional pattern with name "QualityDescriptor" which can be associated to documents which fulfill a certain quality standard. It could be used as additional filter for a search.

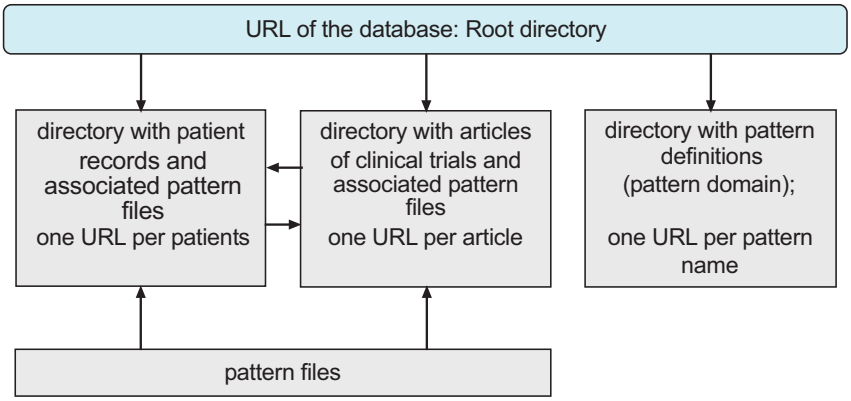

Fig. (2). Proposed initial directory structure of the database. All relevant data are accessible *via* URL. There are links of articles to patient records and backwards, and links of the pattern files to the associated text.

The Search

For a pattern search the user must provide at least the pattern name and the feature vector. In case of patterns with short representation this can be done directly by typing, else indirectly by sending a file with all information to the database. Generation of such a file can be done e.g. by software which is connected with diagnostic devices, or software of laboratories which make blood analysis, or software of radiologists etc. (see examples). More complex search requests are possible using a regular expression, e.g. for AND combination of pattern search with conventional text search.

Combination of Patterns, Multidimensional Search

It would be also possible to combine different patterns for a search. At this the weight for determination of the search result order of every pattern could be predefined by an additional number, e.g. as relative percentage.

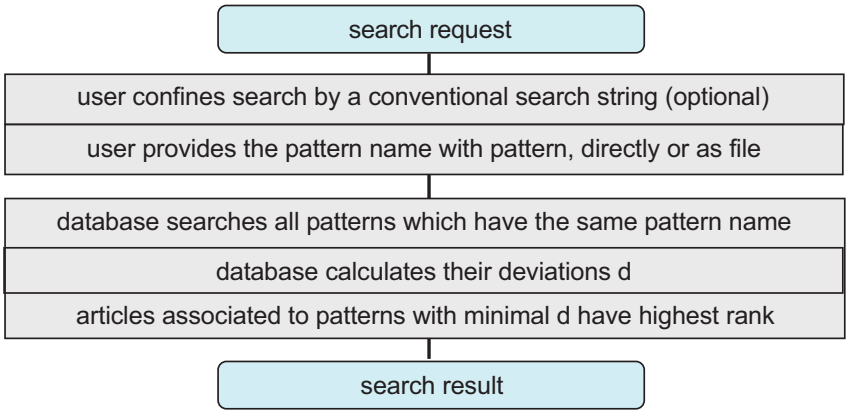

Fig. (3). The search process.

Immediate Local Statistics

Connected with search is the possibility for "local statistics" which is done "near" the current patient: Well-structured articles with clinical records should include not only diagnostic data but also treatment and the patient's rating as searchable patterns. These are fully machine-readable. So it would be possible to collect all articles whose diagnostic patterns are similar[3] to those of an actual patient and calculate at once the mean rating of former "similar" patients in case of this or that treatment. This can help also a general practitioners who wants to decide to which specialist he should send the patient. Today selection of a specialist can be already a preliminary decision for therapy.

Predictions from an Up to Date Model

A local statistics is not difficult to interpret, moreover it has the advantage that it can be quickly calculated. But especially if there are only a few patient records "near" the actual patient, the result could be imprecise due to coincidental fluctuations and it can be more accurate [9] to consider all patients records with the same kind of pattern like the current patient for immediate regression analysis or another appropriate precalculation, and make predictions from this. This could be even save computational time, because new precalculation is only necessary after new pattern of this kind have been uploaded.

DISCUSSION

Feasibility, Possible Problems and Solutions

Investment Costs

In 1991 arXiv [22] has been founded. It is an archive for e-prints of scientific papers in the fields of physics, mathematics, computer science and quantitative biology which can be accessed *via* the Internet. In many fields of mathematics and physics, almost all scientific papers are placed on the arXiv. As of June 2007, arXiv.org contains over 423,000 e-prints, with roughly four thousand new e-prints added every month. The costs of it have been estimated in [23]:

"In combination with system maintenance and upgrades and managerial and administrative support, costs of arXiv operation should average about $200,000 annually".

This should only illustrate the possible efficiency of such archives. The authors do most work, and they do it without payment.

In comparison to arXiv a database with patterns search option needs additional manpower and we expect that complexity of requirements increases in medicine. In the course of time can emerge special needs, e.g. conversion and recalculation of patterns from original data, construction of interfaces for web services. Parallely to this traffic would increase. Then allowance of discreet advertisements and other activities[4] could secure solid finances. On these condition private funding is possible. Public funding, however, could guarantee free access to the database. This should be discussed.

Inappropriate Uploads

All reality conform[5] documentations of patient histories and of clinical trials are welcome. We hope that, like in natural science, also in medicine there will be many authors who want to share their knowledge and experiences, and that self-control of these authors works well, too.

The database's personnel could control formal requirements and rough content. Additionally all registered users have the possibility to rate any upload. Later the user who downloads information may decide individually[6] whether rating and other criteria, e.g. author's specialization, should be relevant for a search, or whether certain groups of authors should be not taken into consideration.

Quality of Pattern Definitions

Language is necessary for initial definition of new patterns. Suitable patterns can be defined without room for misinterpretation. Because definition is done once only in the pattern domain, the patterns remain unambiguous. So they

[3] "Similarity" means that their distance d is smaller than a given maximum.

[4] For example the database could support websites which initiate new contacts between doctors and patients, for a small fee, like ebay™ between seller and buyer, or initiate contacts in self-help groups, hopefully without fee. We can only mention this here, because more would exceed the scope of this article.

[5] May be that some authors tend to sugarcoat the outcomes of own therapies. Of course this would be inappropriate, and the database cannot avoid this, but its open structure helps to detect this in the long run. Every author is responsible for his/her contribution.

[6] These individual preferences could be stored as file to be quickly retrievable.

represent well-defined machine-readable medical information. This aspect could contribute to the Unified Medical Language System (UMLS) [24].

Reliability of Pattern Definitions

Redefinition of a pattern with given name should be avoided, instead of this the new definition should be associated to a new pattern name within the same group, e.g. by simply appending an increasing number to the initial pattern name. If original data have been uploaded, calculation of new feature vectors with new associated pattern names would be possible even retroactively.

Integrity

As shown in Fig. (**2**), all contents of the database are accessible *via* web URLs. The initial structure is as simple and robust as possible. If there is a dead link, the rest remains intact. The database is fully open only for read. Writing of patient records, comments and articles to the database can be done only additively *via* upload. Deletion is possible only by the administrator.

At least one remote mirror of the database is necessary.

Complexity

It is advisable to think about possible barriers concerning the computational complexity of the project.

Complexity of Data Storage and Update

The storage space complexity seems to be no great problem because even the repeated backup of the total web has been done [25].

Time Complexity

The search time mainly depends on the dimensionality of the pattern representation, on the number of comparable patterns on the database which are only those with the same pattern name[7] and on further confinement of the search by a conventional search string. Due to this preselection the subset of concerned patterns is usually relatively small, so that a quick search is possible. Quick search is also possible in case of many concerned patterns, if their dimensionality is small enough for an appropriate tree structure [26-30]. Only if very many high dimensional patterns with the same pattern name are stored on the database, and if the search is not enough confined, the search time can become critical. In this case we could use parallelization techniques [31]. If we accept small errors, this could be combined with dimension reduction [32] and approximation methods [33-35].

Privacy

The name of the author must be published. The names of patients are not published, and the complete birth date is invisible, only the birth year. The patient history can be only uploaded, if the patient explicitly agrees to that. On that condition the patient gets a secret number which will be associated to all future documentations which concern him. Using this number and a password also the patient can look in the database for these documentations and rate them. Later the investigator can decide how to use this additional information. Communication protocols which provide both anonymity and personal feedback have been proposed [36].

INNOVATIVE ASPECTS

There are already numerous medical databases which contain collections of patient histories, usually private and only used for one or a few clinical trials. There are also larger medical databases like the Duke Databank for Cardiovascular Diseases [37] and the ARAMIS Project [38] which led to decades of valuable clinical publications and show the benefits of open data collections. But all these databases are confined to their special application and the number of diagnostic patterns is very limited, because these are defined centrally by a few persons, e.g. by some developers.

The here described database organization overcomes this difficulty: The diagnostic and all other numerical patterns are *not* defined by the database's personnel, but decentrally by the "pattern domain owners". Together they have much more working capacity[8] and can develop and describe an increasing number of valuable concepts for reproducible conversion of medical reality into appropriate feature vectors which represent the patterns.

The database's personnel concentrate on development of well-defined and efficient ways for handling of these numerical forms, so that the usual problems like similarity search, comparison, statistics and modelling can be solved quickly. So the database can be used universally for handling of *all* appropriately defined patterns. Language independent pattern search is one important application with obvious advantages. It can be easily combined with conventional text search.

FROM ORIGINAL DATA TO SEARCHABLE PATTERNS

Hints for Definition

It is not difficult to define patterns according to current research - the feature vectors can contain all necessary data which are measured in clinical trials. The associated patient records can be uploaded and the results are directly accessible and comparable. For identification of diseases among others a pattern defined according to the International Classification of Diseases (ICD) can be used. Parallely to this it would be advantageous to search for pattern definitions which systematically map subjective and physical similarities of symptoms to similarities of the feature vectors. One of the first steps would be to define appropriate curved coordinate systems for the constituents of the human body.

EXAMPLES

First Example, Initial Considerations

The initial considerations which lead to development of this idea arose from the field of orthodontics, which is concerned with the study and treatment of malpositioned teeth and

[7] These would be accessible at once using an alphabetically sorted pattern name index. Each entry of this index can point to cached collections of patterns with the same pattern name. Then these can be compared as quickly as possible.

[8] Recall the immense work done by owners of Internet domain names.

the control and modification of facial growth. Cephalometrics is done for treatment planning [39, 40]. At this lateral skull radiographs are taken under standardized conditions and measured (Fig. **4**). The results can be used for building feature vectors. Using these data improves prognosis of skeletal growth.

Advanced orthodontic techniques use devices which directly digitize the three dimensional tooth positions [41] (Fig. **5**). From this data coordinate system independent feature vectors can be calculated for treatment planning.

We noticed, that the approach can be generalized. Feature extraction of diagnostic findings is also possible in other areas of medicine. Often such findings are the basis of severe decisions. The following examples illustrate this.

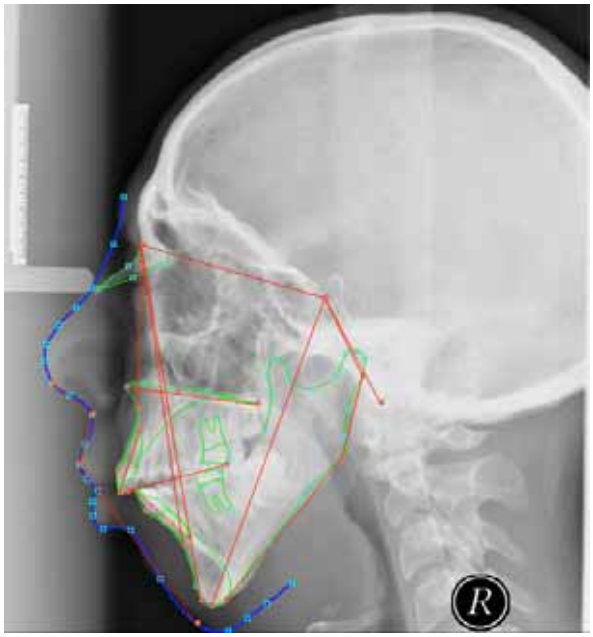

Fig. (4). Cephalometric analysis; the angles and distances measured on the lateral skull radiographs (cephalograms) can be used for building a feature vector of this profile.

Second Example, Nontrivial Application and Prototype

Sometimes complex original data can need complex precalculation. If simple self-evident considerations (Fig. **10**) are not enough, an appropriate transformation of pictures, sounds or curves is often the first step for calculation of feature vectors. For example in case of heart sounds a wavelet transformation allows analysis of the signal at different scales and times. Initially it is necessary to select and border

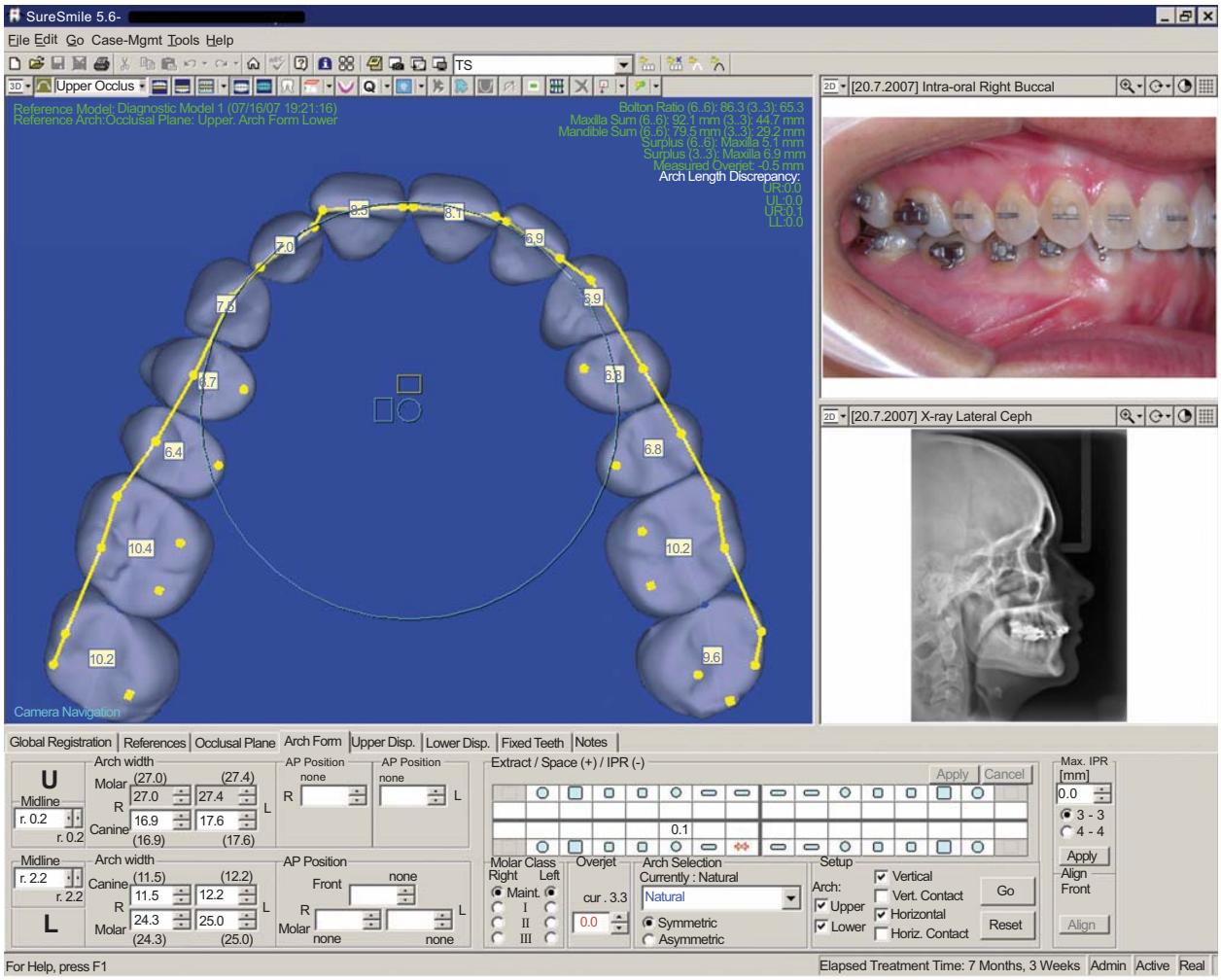

Fig. (5). Measurements on the digitized three-dimensional surface of the dentition. Digitizing device: OraScanner™ (Orametrix, Inc. in Dallas).

accurately a representative period of the sound (Fig. **6a**-**c**). The resulting wavelet coefficients (Fig. **7**) can be used for building the feature vector which represents the pattern.

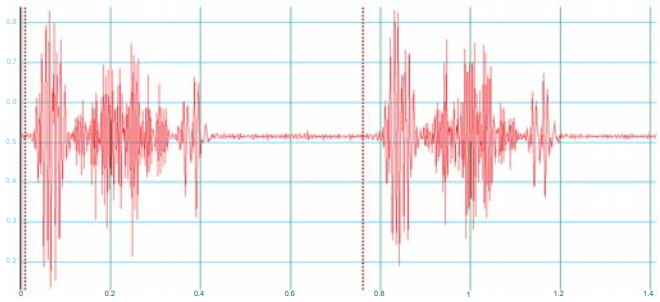

Fig. (6a). Heart sounds in case of aortic valve stenosis; vertical axis: relative amplitude, horizontal axis: time in seconds. The brown dashed lines represent a bordering of the first period which has been set approximatively by the user on the screen.

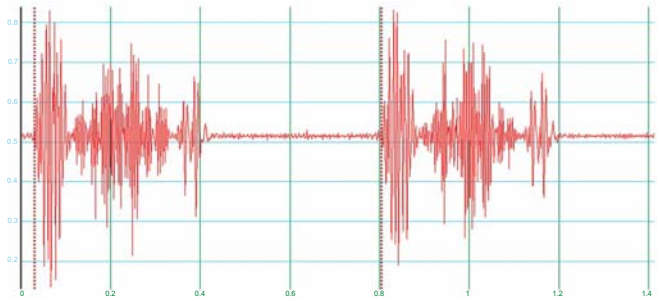

Fig. (6b). A catching algorithm is applied which reproducibly refines the bordering of Fig. (**6a**).

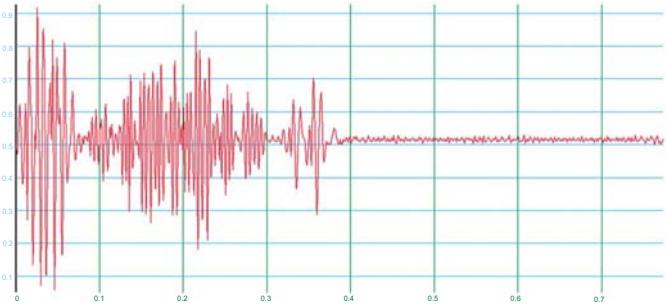

Fig. (6c). The bordered part of Fig. (**6c**) is stretched that exactly one period remains.

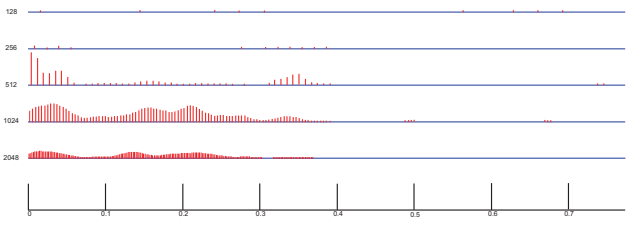

Fig. (7). The sound of Fig. (**6c**) after Daubechies wavelet transform; smoothened absolute values of the transformation coefficients for five different scales. They can be directly used for building a searchable feature vector of this heart sound.

Development and improvement of such calculations requires research. Remembering the variety of useful pattern structures it becomes clear that sharing of the work is necessary. The database cannot generate searchable patterns (Fig. **7**), but it can store, compare and rank them (Fig. **8**).

In the next chapter we show that even complex original data like MRI scans can lead to very compact patterns.

Third example, Case Report

A common question to the database will be: Should we operate, and if yes, which operation has best results. If the operation has severe consequences, there must be a good justification for it. For example Fig. (**9a**) shows the MRI of a 2 weeks old osteoporotic[9] compression fracture in the area of maximal kyphosis of the thoracic spine. The treating surgeon was a specialist in doing spinal fusions and recommended a dorsoventral spondylodesis. The patient trusted and accepted. After the extensive operation the patient read about kyphoplasty [42] and heard from experts that this minimalinvasive method would have been adequate in his case. So he got the impression that the operation trauma (lateral thoracotomy, muscle dissection), which leads to chronic pain, has not been necessary. The surgeon, however, dislikes kyphoplasty [43] and remains committed to his operation (Fig. **9b**). Obviously medical treatment dramatically depends on the experience[10] of the chosen doctor.

This is no good situation. A comprehensive and uniform source of information like the proposed one could help and avoid dissent. If there would have been a good searchable database, before recommendation the surgeon could have looked for similar[11] fractures in the database and could have asked for chances of success in case of conservative therapy, kyphoplasty, minimally invasive endoscopic surgery [44], other therapies, or dorsoventral spondylodesis with lateral thoracotomy. If the latter would have been best, he could recommend *and* justify a large operation, there would not be any problem. If another possibility would have been better, he could early enough recommend another treatment and avoid a large unnecessary operation trauma with all consequences.

PERSPECTIVE

If accepted, the database would have significant influence both on research and on clinical practice. Using the database future researcher could systematically look for certain features of the human body, of medical decisions and of

[9]The male patient was only 44 years old and an endocrine reason for his osteoporosis has not been found. But there was a particularity: Since the age of 18 he performed intensive daily ergometer training, and kept underweight, the BMI has been 17-18. When looking at literature at the age of 18 he got the impression that this lifestyle is good for health and kept it up because he felt fit. The osteoporotic fracture was very surprising for him and his environment. After noticing the danger for bones, he corrected weight and sport. Ergometer training was done less intensive and partially replaced by strength training. Within 16 months the DXA T-score at L2 increased from -4,6 to -3,4, the bone density increased about 20% within this time!

[10]Of course global experience is too large to be captured by a single person. But we can make it better accessible by collecting it and providing an adequate interactive possibility to find its part which is relevant for the current situation.

[11]With similar (v,c,d,n,t) as defined in Fig. (**10**).

Fig. (8). Exemplary output of our database prototype. The uploaded pattern represents the heart sound in case of aortic valve stenosis after wavelet transformation as shown in Fig. (**7**). Links to articles with most similar stored patterns are listened first. The links are accompanied by structured information for test purposes. The distance d quantifies the deviation to the uploaded pattern, the first link points to an article with the same sound like the uploaded one, therefore the distance is zero.

lifestyle which later correlate significantly with certain trends of health. The doctor could select these features and search with them, depending on the clinical picture, and look for decisions which later correlate with best possible health.

If there are no adequate symptoms but relevant risks, preventive selection of diagnostics according to statistics can be adequate. An example:

Prophylactic MRI Scans

We have seen the great significance of MRI scans. Because they seem to be innocuous, we recommend more extensive usage of this possibility also for prophylaxis:

Nearly all of us have lost an affiliated person due to cancer which would have been detectable in an early state by MRI. Cancer is so frequent and so painful that we suggest as prophylaxis periodically[12] standardized MRI scans of all interested people. At least scans of imperiled tissues should be done at an age, in which these frequently lead to detection of a serious disease[13]. Three-dimensional imaging is possible. From the most significant scans feature extraction could be done. The resulting feature vectors could be stored as searchable patterns in the database. From this we could systematically learn about feature changes, which later correlate with serious diseases. This would lead to a well-founded basis for efficient MRI prophylaxis.

Of course such scans should be combined with other measurements, e.g. blood tests like PSA, if statistically meaningful. After establishment of the database we expect competition of diagnostic methods - the most meaningful methods can be easier recognized and selected.

Advanced Possibilities for Evaluation

Besides efficient search there are additional possibilities. Immediate individual statistics and regression analysis already have been mentioned. Because the patterns are machine-readable, they could be evaluated automatically by software, e.g. for conversion, modelling - it would exceed the scope of this article to deepen this here.

[12]Of course health insurance can only pay a limited number of scans. But the patient should know about their high detection rate, so that he/she could knowingly decide where to invest money.

[13]Not only cancer, Fig. (**6a**) for example shows old osteoporotic fractures. Of course we know that it is difficult for the radiologist to evaluate a great number of scans of a patient which has no symptoms. Therefore should be discussed that the radiologist is freed from liability in case of underdiagnose. Else there is especially in the initial state, without experience from an existing large database, the danger of many false positive diagnostics.

Fig. (9). (**a**) osteoporotic compression fracture (**b**) chosen therapy: dorsoventral spondylodesis.

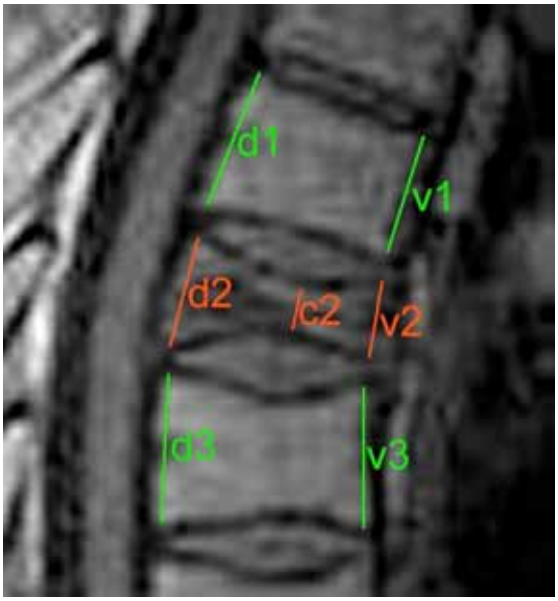

Fig. (10). Drawn are the lines with lengths d1, d2, d3, c2, v1, v2, v3 which are used for classification of the vertebral compression fracture and for calculation of three important numbers (v,c,d) which provide decision-relevant information about the geometry of the fracture. They show the relative remainder of the vertebral body ventral (v), central (c) and dorsal (d). Let d1, d2, d3, c2, v1, v2, v3 denote the scalar lengths of the lines drawn in Fig. (**10**). Then we can calculate the numbers v, c, d as follows: v=2*v2/(v1+v3), c=4*c2/(v1+v3+d1+d3), d=2*d2/(d1+d3). Additionally the number n of the vertebra and a representative measurement of bone density t like the DXA T-score may be important, so that we could define the 5-dimensional feature vector (v,c,d,n,t) for a vertebral compression fracture. For a search it is preferable, but not necessary to know all 5 values. If, for example, someone has a compression fracture and only c and n are known, it is possible to search in the database for all vertebral fractures with similar c and n.

CONCLUSIONS

Up to now a great part of medical experience gets lost. But it would be technically feasible to collect anonymously medical documentations on a voluntary basis in an increasing global database and to search in it efficiently using diagnostic patterns. It could be also used for immediate calculation of up to date prediction models. Associated pattern names and feature vectors induce a systematic and arbitrarily expandable approach. Funding is necessary for good software and hardware support of the database, the definition of appropriate pattern names and feature vectors can be shared among medical institutions and further interested parties, e.g. manufacturer of diagnostic devices. The interface can be designed to allow comfortable upload as well as comfortable download of publications and patient records. We hope to start a constructive discussion which at last leads to realization of the project. All creative suggestions to this are welcome.

SUMMARY POINTS

What is Usual Up to Now:

- Databases with collections of patient histories are standard for documentation of clinical trials and medical studies. They are separated, specialized and often private.
- All these databases are confined to their special application and the number of diagnostic patterns is very limited, because these are defined centrally by a few persons, e.g. by some developers.

What the Proposed Database Offers:

- The proposed database has an universal open and worldwide accessible interface. New anonymous patient records can be added interactively, if wished linked together with associated documentations of clinical trials.
- At this the number of allowed diagnostic and other well-defined numerical, machine-readable patterns is not limited. Their definition is done decentrally by "pattern domain owners", e.g. researchers, manufacturer of diagnostic devices, which have much more working capacity than the database's personnel. Any owner of an Internet domain name is automatically also owner of a pattern domain which starts with the same name.
- The database allows numerical search for these patterns, if wished combined with conventional text search. Furthermore it provides standard algorithms for their numerical evaluation, like statistics and modelling. Periodically calculation of up to date prediction models is possible.

REFERENCES

[1] Snow PB, Smith DS, Catalona WJ. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. J Urol 1994; 152: 1923-6.

[2] Partin AW, Kattan MW, Subong EN, *et al.* Combination of prostate-specific antigen, clinical stage, and Gleason score to predict

pathological stage of localized prostate cancer. A multi-institutional update. JAMA 1997; 277(18): 1445-51.
[3] Ross PL, Scardino PT, Kattan MW. A catalog of prostate cancer nomograms. J Urol 2001; 165: 1562-8.
[4] Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. J Natl Cancer Inst 1998; 90: 766-71.
[5] Stephenson AJ, Scardino PT, Eastham JA, *et al*. Preoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. J Natl Cancer Inst 2006; 98: 715-17.
[6] Kattan MW, Wheeler TM, Scardino PT. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. J Clin Oncol 1999; 17: 1499-507.
[7] Stephenson AJ, Scardino PT, Eastham JA, *et al*. Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. J Clin Oncol 2005; 23(28): 7005-12.
[8] Kattan MW. Nomograms. Introduction. Semin Urol Oncol. 2002; 20(2): 79-81.
[9] Kattan MW, Leung DHY, Brennan MF. Postoperative Nomogram for 12-Year Sarcoma-Specific Death. J Clin Oncol 2002; 20(3): 791-96.
[10] Wong SL, Kattan MW, McMasters KM, Coit DG. A Nomogram That Predicts the Presence of Sentinel Node Metastasis in Melanoma With Better Discrimination Than the American Joint Committee on Cancer Staging System. Ann Surg Oncol 2005; 12(4): 282-8.
[11] Kattan MW, Karpeh MS, Mazumdar M, Brennan MF. Postoperative Nomogram for Disease-Specific Survival After an R0 Resection for Gastric Carcinoma. J Clin Oncol 2003; 21: 3647-50.
[12] Eggener SE, Yossepowitch O, Pettus JA, Snyder ME, Motzer RJ, Russo P. Renal Cell Carcinoma Recurrence After Nephrectomy for Localized Disease: Predicting Survival From Time of Recurrence. J Clin Oncol 2006; 24(19): 3101-6.
[13] Specht MC, Kattan MW, Gonen M, Fey J, Van Zee KJ. Predicting nonsentinel node status after positive sentinel lymph biopsy for breast cancer: clinicians versus nomogram. Ann Surg Oncol 2005; 12: 654-9.
[14] International Bladder Cancer Nomogram Consortium. Postoperative Nomogram Predicting Risk of Recurrence After Radical Cystectomy for Bladder Cancer. J Clin Oncol 2006; 24(24): 3967-72.
[15] Bach PB, Elkin EB, Pastorino U, *et al*. Benchmarking Lung Cancer Mortality Rates in Current and Former Smokers. Chest 2004; 126(6): 1742-9.
[16] Brennan MF, Kattan MW, Klimstra D, Conlon K. Prognostic nomogram for patients undergoing resection for adenocarcinoma of the pancreas. Ann Surg 2004; 240: 1-6.
[17] Böhm C, Berchtold S, Keim DA. Searching in High-Dimensional Spaces: Index Structures for Improving the Performance of Multimedia Databases. ACM Computing Surveys 2001; 33(3): 322-73.
[18] W Orthuber. Universal pattern search on the web. http://www.orthuber.com/wpa.htm, since 2006-06-02.
[19] Wikipedia, Metric (mathematics) http://en.wikipedia.org/wiki/Metric_(mathematics), visited 2008_02_10
[20] W3C. Naming and Addressing. http://www.w3.org/Addressing/, viewed 2007-07-24.
[21] MRInstitute leading the way to electronic medical records, http://www.medrecinst.com/, visited 2007.
[22] http://arxiv.org/. A fully automated electronic archive and distribution server for research papers. since 1991.
[23] Hickerson HT. Project Euclid and the ArXiv: Complimentary and Contrasting Elements for Sustainability. An edited version of remarks presented at the Workshop on Sustainable Models for University-Based Scholarly Publishing, conducted at Columbia University on June 1 2004.
[24] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993; 32(4): 281-91.
[25] Internet Archive Wayback Machine http://www.archive.org/ index.php, founded in 1996.
[26] Beckmann N, Kriegel HP, Schneider R, Seeger B. The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. SIGMOD Conference 1990; 322-31.
[27] Berchtold S, Bohm C, Kriegel HP. The Pyramid-Tree: Breaking the Curse of Dimensionality. SIGMOD Conference 1998; 142-53.
[28] Berchtold S, Keim DA, Kriegel HP. The X-tree: An Index Structure for High-Dimensional Data. VLDB 1996; 28-39.
[29] Beckmann N, Kriegel HP, Schneider R, Seeger B. The R*-tree: an efficient and robust access method for points and rectangles. Proceedings of the 1990 ACM SIGMOD international conference on Management of data 1990; 322-31.
[30] Berchtold S, Böhm C, Jagadish HV, Kriegel HP, Sander J. Independent Quantization: An Index Compression Technique for High-Dimensional Data Spaces. In Proc 16th Int Conf on Data Engineering 2000.
[31] Berchtold S, Böhm C, Braunmüller B, Keim DA, Kriegel HP. Fast parallel similarity search in multimedia databases. Proceedings of the 1997 ACM SIGMOD international conference on Management of data 1997; 1-12.
[32] Fodor IK. A survey of dimension reduction techniques. US DOE Office of Scientific and Technical Information 2002.
[33] Weber R. Similarity Search in High-Dimensional Vector Spaces. Theses to databases and information systems 2001; 74 ISBN 3-89838-474-8.
[34] Nene SA, Nayar SK. A simple algorithm for nearest neighbor search in high dimensions. IEEE Trans Pattern Anal Mach Intell 1997; 19(9): 989-1003.
[35] Brin S. Near Neighbor Search in Large Metric Spaces. Proceedings of the 21th International Conference on Very Large Data Bases 1995; 574-84.
[36] Luttenberger N, Kollek R, Reischl J, Stürzebecher CS. Design of Individual Donor Feedback Processes in Biobank Research. In Proceedings of GI Jahrestagung, LNI 93, Springer 2006; 722-28.
[37] Duke University Medical Center, http://www.duke.edu/
[38] McShane DJ, Fries JF. The chronic disease data bank--the ARAMIS experience. Proceedings of the. IEEE 1988; 76(6): 672-9.
[39] Athanasiou E. Orthodontic Cephalometry. London: Mosby-Wolfe, 1995.
[40] Schutyser F, Hausamen AC, Swennen JE. Three-dimensional cephalometry, G. R. J. A color atlas and manual. Berlin Heidelberg New York: Springer; 2006.
[41] Orametrix®, http://www.orametrix.com/, visited 2008_02_10.
[42] Kasperk C, Hillmeier J, Nöldge G, *et al*. Kyphoplastie-Konzept zur Behandlung schmerzhafter Wirbelkörperbrüche. Dtsch Arztebl 2003; 100(25): 1748-52.
[43] Hopf C, Heeckt, H. Kyphoplastie – Konzept zur Behandlung schmerzhafter Wirbelkörperbrüche: Kommentierung notwendig. Dtsch Ärztebl 2003; 100(44): 2882-3.
[44] Beisse R, Potulski M, Bühren V. Endoscopic surgery of the thoracic and. lumbar spine-A report on 500 treatments. Osteosynthesis Trauma Care 2003; 4:196-205.

Received: October 31, 2007 Revised: January 25, 2008 Accepted: February 23, 2008