Medical Informatics in a United and Healthy Europe
K.-P. Adlassnig et al. (Eds.)
IOS Press, 2009

doi:10.3233/978-1-60750-044-5-584

A Searchable Patient Record Database for Decision Support

Wolfgang ORTHUBER [1], Thorsten SOMMER
Department of Orthodontics,
Universitätsklinikum Schleswig-Holstein Campus Kiel, Germany

Abstract. We describe a searchable patient record database for decision support. It contains medical histories of real but pseudonymous patients with patterns of diagnosis, chosen treatment, and outcome. To be searchable, the patterns contain a feature vector (for similarity search by calculating distances) and a globally unique "pattern name" which identifies the kind of data which are represented by the feature vector. Patterns with the same pattern name are directly comparable; they represent the same kind of data. For pattern selection the database provides a growing well-structured list of initial diagnoses with associated input masks. Procedure: The doctor can assume that the database contains patients similar to the current patient if he finds his initial diagnosis in the list. Clicking on it opens an associated input mask which requests specific further data for finer differentiation. After input a searchable pattern group is built from the provided data, and used to search for histories of patients with similar fine diagnostics, and for the most successful treatment decisions at these patients. This information can be very valuable for deciding the treatment of the current patient. Because the database can collect patient histories from all countries, in the long run this could open access to a wealth of experience which by far exceeds the capacity of a today's doctor.

Keywords. similarity search, patterns, pattern names, feature vectors, patient record database, decision support

1. Introduction

Modern medical knowledge is growing rapidly, only a small part of it can be captured by a human doctor. 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.

1.1. Background

For decision support computer-aided applications [1] and prediction tools have been developed, for example neural networks [2], probability tables [3] and nomograms [4, 5]. These use the experience from thousands of patients and are of particular importance in case of clinical decisions with serious consequences. Today there are so many models, that for some situations selection becomes difficult.

[1] Corresponding Author: Mathematician, Orthodontist, Arnold Heller Str. 16, 24105 Kiel, Germany; E-mail: orthuber@kfo-zmk.uni-kiel.de.

All models are derived from (expert knowledge from) collections of patient histories. Meanwhile the web allows the more direct way: To store all these and further documentations in a global database in a way which allows efficient information retrieval for decision support.

1.2. Goal

From this the goal of this paper is derived: To describe the design of a global database for up to date decision support. Concretely the database should provide for the patient's diagnostic results:

- possible therapies
- their long term consequences with probabilities of failure and success

2. Material and Methods

2.1. Organization of the Database

The database is designed for efficient collection of patient records. It accepts uploads in all browser-readable formats. Data which are uploaded to the database get a time stamp and cannot be deleted. Due to this WORM (write once read many times) feature it can be used for conclusive documentation. Doctors have always access to their own uploads, and access in case of emergency. This is recorded in a log file. Further access and permission to access is fully controlled by the patient using a secure key, e.g., a card. The patient may decide to publish nothing (default) or only a part, e.g., machine readable patterns. Participation is voluntary. Every patient who wants to participate gets a unique pseudonym which is also the name of his directory in which his data are stored. Uploads to the database are stored sequentially, every upload gets (as "record") a new URL. So every patient has a main URL and all data are (in case of permission) accessible via URL (Figure 1).

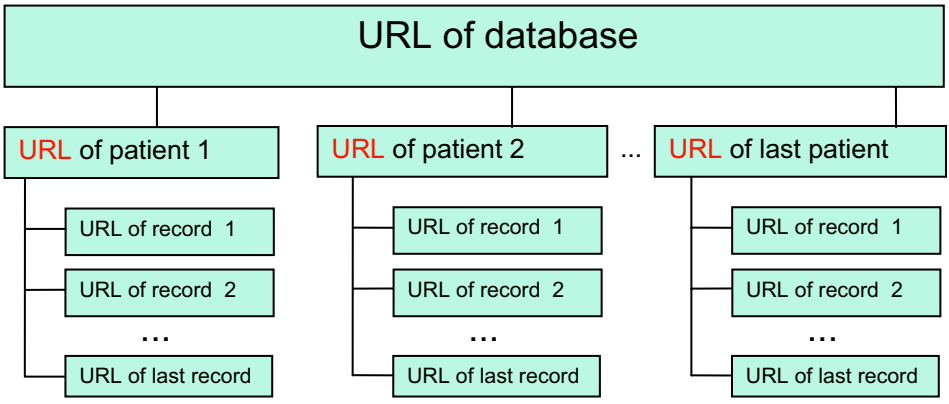

Figure 1. All data are stored pseudonymously and accessible via URL

For decision support the uploaded records must contain all decision-relevant information (diagnostic data, chosen therapy and therapeutic results) in searchable,

precise and machine-readable form. A data structure called pattern fulfills the requirements, its main components are a pattern name and a feature vector (Figure 2a).

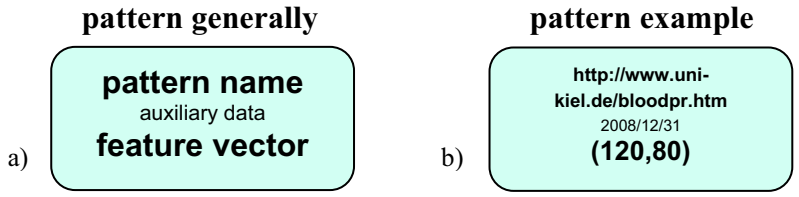

Figure 2. a) A searchable *pattern* according to our definition. **b)** Exemplary content of a pattern. The pattern name http://www.uni-kiel.de/bloodpr.htm is a http URI which internationally uniquely identifies the meaning of the feature vector *(120,80)* as measurement result of blood pressure in the form (S,D) in which S = systolic blood pressure in mm Hg and D = diastolic blood pressure in mm Hg.

The pattern name identifies the meaning of the numerical content (feature vector) internationally uniquely. So it is clear which kind of original data is represented. Further auxiliary data provide additional information, e.g., date, links to URLs which are associated with this pattern. Feature vectors of patterns with the same pattern name (patterns "of the same kind") are directly comparable using an associated metric [6–8]. The result of such a comparison is a distance $d \geq 0$ which quantifies similarity. The smaller the distance, the more similar are the patterns. Patterns with zero distance have identical feature vectors.

The pattern name is a http URI [9, 10] which (like a URL) permanently refers to a unique, permanent web address and which differs if and only if the web address differs. So it is a unique name and also a unique reference. It points to a linking file (http://www.uni-kiel.de/bloodpr.htm in Figure 2b) which points to all defining and further associated information. **This guarantees also that there is exactly one determining definition (anchor) for this name, and indirectly well defined task sharing among all domain name owners for definition of patterns.**

2.2. How to Find in the Database Decision-Relevant Information

To keep overview a branching strategy is necessary. For this the database provides an increasing well-structured list of terms which describe initial (rough) diagnoses and/or symptoms. It can be partially derived from current classifications, e.g., ICD [11], SNOMED CT [12]. First the doctor has to find a term in this list which best describes the initial diagnosis or the main symptom. If the doctor finds an appropriate term, he can assume that the database contains related patient histories, and click on it. The database opens a concise input mask which asks for finer decision-relevant diagnostic data, well-adapted to the selected term, e.g., specific laboratory data and results of measurements on the patient. An appropriate group of patterns with weighting coefficients is associated with every input mask. Two pattern groups X and Y are comparable, if for every pattern in group X there is exactly one corresponding pattern with the same pattern name in group Y. The distance between two pattern groups is the sum of distances between the corresponding patterns, respectively multiplied by the associated weighting coefficients.

3. Results

After the doctor has completed the input mask, an associated pattern group A is calculated from the provided data. Using an index table the database quickly selects all patients whose records contain a comparable pattern group according to chronological rules which are specified in the input mask. Then it calculates their similarity by calculating the distances of the found pattern groups to the pattern group A. In the search result the URLs of found patients and associated clinical records with minimal distance are listed first, i.e., medical histories of those patients are listed first whose data are most similar to the data which the doctor has provided in the input mask. It is also possible to use a regular expression to specify additional conditions, e.g., ranges of certain components in the feature vectors.

It depends on the size of the database whether patients can be found who are "similar enough" for decision support. The doctor can check the data of the found patients. In case of sufficient similarity the information about most successful treatment decisions can be very valuable for deciding the treatment of the current patient. One may define the term "sufficient similarity" by providing a maximal distance. If there are enough patients with distance below this maximum, a local statistics among these "similar" patients can provide probabilities of success in case of this or that treatment decision, as desired in section 1.2.

4. Discussion

It is not possible to simulate the global database in an experiment, the above results are based on conclusion by analogy: (1) The domain name system induced a world-wide task sharing for creation of the web and has been very successful. It is plausible and efficient to use this already existing system also as basis of world-wide task sharing for definition of patterns. (2) Similarity search in metric spaces is well explored [13], a nontrivial local prototype has been demonstrated (similarity search of heart sounds [6]). For larger amounts of data we made a small well-defined experiment to get a hint about the necessary time for comparison of patterns using a C++ compiler on a single PC (2,1 GHz Pentium). The time for calculating 1 million weighted Euclidean distances of vectors with dimensionality 10 in double accuracy was between 0.20 and 0.21 seconds. Such calculation would be only necessary if similarity search is done within patterns with altogether 1 million 10 dimensional feature vectors with serial comparison. Depending on implementation, disk access would require additional time. To minimize the time for comparison the total dimensionality (the sum of dimensions of all feature vectors) of the searched pattern group should be minimized. This should be also done because the sparsity increases exponentially with the dimensionality in case of a constant amount of data, with points tending to become equidistant from one another ("Curse of Dimensionality" [14, 15]).

The most important problem seems to be to get enough uploads of patient records with machine-readable searchable patterns, so that the database becomes attractive. It would be necessary that the vendors of patient record software provide an interface for export of clinical records in browser readable and pseudonymous form. These must already contain searchable patterns. Therefore the database should provide the option that patterns produced from input masks are (besides usage for search) also sent as file back to the user. So every user can produce searchable patterns. These could be

imported by patient record software before upload. So important decision-relevant medical information can be generated in reproducible way and stored in internationally well-defined and machine-readable form.

5. Conclusion

It is possible to design a large patient record database which allows similarity search for objectifiable clinical findings. The doctor enters most important diagnostic data and the database searches for histories of patients with similar fine diagnostics, and for treatment decisions which have been most successful at these patients. Because of the machine readability of decision-relevant data there are enhanced possibilities, e.g., local statistics, immediate calculation of prediction models, further mathematical modeling.

References

[1] Medexter. Providing solutions for medical decision support, www.medexter.com/index.php.
[2] P.B. Snow, D.S. Smith, W.J. Catalona. (1994) Artificial neural networks in the diagnosis and prognosis of prostate cancer: A pilot study. *Journal of Urology* 152(5 II):1923–1926.
[3] Partin, A.W., Subong, E.N., Walsh, P. et al. (1997) Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathological stage of localized prostate cancer. A multi-institutional update. *Journal of the American Medical Association* 277(18):1445–1451.
[4] Stephenson, A.J., Scardino, P.T., Eastham, J.A. et al. (2005) Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *Journal of Clinical Oncology* 23(28):7005–7012.
[5] Kattan, M.W. (2002) Nomograms. Introduction. *Seminars in Urologic Oncology* 20(2):79–81.
[6] Orthuber, W., Fiedler, G., Kattan, M., Sommer, T., Fischer-Brandies, H. (2008) Design of a global medical database which is searchable by human diagnostic patterns. *The Open Medical Informatics Journal* 2:21–31.
[7] Orthuber, W. Universal pattern search on the web, www.orthuber.com/wpa.htm, since 2 Jun 2006.
[8] Wikipedia. Metric (mathematics), en.wikipedia.org/wiki/Metric_(mathematics).
[9] W3C. Naming and Addressing, www.w3.org/Addressing/.
[10] C. Bizer, R. Cyganiak, T. Heath. How to Publish Linked Data on the Web. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.
[11] WHO. *International Classification of Diseases*, www.who.int/classifications/icd/en/.
[12] International Health Terminology Standards Development Organisation. *SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms)*, www.ihtsdo.org/snomed-ct/.
[13] Zezula, P., Amato, G., Dohnal, V., Batko, M. (2005) *Similarity Search – The Metric Space Approach*. Springer, Berlin.
[14] Indyk, P., Motwani, R. (1998) Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual Symposium on the Theory of Computing*, ACM, New York, 604–613.
[15] Berchtold, S., Böhm, C., Kriegel, H.-P. (1998) The pyramid-technique: Breaking the curse of dimensionality. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, ACM Press, 142–153.