Exploring Complexity in Health: An Interdisciplinary Systems Approach
A. Hoerbst et al. (Eds.)

doi:10.3233/978-1-61499-678-1-257

Collection of Medical Original Data with Search Engine for Decision Support

Wolfgang ORTHUBER[1]
University Clinic Schleswig-Holstein, 24105 Kiel / Germany

Abstract. Medicine is becoming more and more complex and humans can capture total medical knowledge only partially. For specific access a high resolution search engine is demonstrated, which allows besides conventional text search also search of precise quantitative data of medical findings, therapies and results. Users can define metric spaces ("Domain Spaces", DSs) with all searchable quantitative data ("Domain Vectors", DSs). An implementation of the search engine is online in http://numericsearch.com . In future medicine the doctor could make first a rough diagnosis and check which fine diagnostics (quantitative data) colleagues had collected in such a case. Then the doctor decides about fine diagnostics and results are sent (half automatically) to the search engine which filters a group of patients which best fits to these data. In this specific group variable therapies can be checked with associated therapeutic results, like in an individual scientific study for the current patient. The statistical (anonymous) results could be used for specific decision support. Reversely the therapeutic decision (in the best case with later results) could be used to enhance the collection of precise pseudonymous medical original data which is used for better and better statistical (anonymous) search results.

Keywords. High Resolution Description, Numeric Search, Metric Space, Domain Space, DS, Domain Vector, DV, Feature Vector, Interoperability, Standardization

1. Introduction

We want to emphasize that it is not possible to describe all important details of this comprehensive approach in a 5 page article. We have to start with efficient handling of precise original data! If you miss details or want more information, please look at [16]!

Medicine is becoming more and more complex and humans can capture total medical knowledge only partially. Therefore merging of medical experience is desirable, so that therapeutic decisions less depend on the chosen doctor.

Current approaches: Helpful can be enhancement of the own knowledge and in certain cases special decision support systems. Already patients use text search engines like Google. Similarly the doctor can do this. Also the IBM Watson system [7] uses natural language. Recently published are results from a text search engine in original data [5]. The system is progressive in comparison to other existing systems. Problems: The transfer of individual (pseudonymous) patient data to external persons is in many countries prohibited due to data protection laws. Free text is not well defined and can be biased. Very important is the unsharpness of textual descriptions and words. Even if we use terms of the well known SNOMED CT nomenclature [2] at the most

[1] Corresponding Author: orthuber@kfo-zmk.uni-kiel.de

differentiated level, we need finer information for decision support. Examples of such terms: "diabetes-nephrosis syndrome", "diabetic glomerulopathy", "diabetic glomerulonephritis", "diffuse type diabetic glomerulosclerosis", "persistent proteinuria associated with type 1 diabetes mellitus" ...

It becomes soon clear, that even the finest level of large nomenclatures is not enough. We need for therapeutic decisions additional information. Even if a doctor looks at a medical picture, e.g. a radiographic image, at last quantitative data (e.g. localizations, distances in a picture, all decision relevant measurements on the patient) are checked to come to a decision.

Appropriate quantitative description is well defined and (in case of precise data basis) much more accurate than the description by words of language (which are also used by current non quantitative nomenclatures). The cardinality of definable quantitative spaces is much greater than that of language vocabulary.

It is obvious that quantitative data are essential for medical decisions. There are nomenclatures for identifying laboratory tests and clinical observations in electronic messaging like the Logical Observation Identifiers Names and Codes (LOINC) [8]. But these cover only a small part of possible decision relevant data (which also include feature extraction results of complex findings) and up to now there is no systematic approach for making quantitative data searchable, though searchable data are essential for efficient decision support. The problem can be solved:

2. Solution

Already in [13] the potential of standardized vectorial representation of medical data in patient records has been described. Here we do not repeat all details. Meanwhile the data model has been extended with new nomenclature, and an online implementation is available in http://numericsearch.com [14]. To make quantitative data searchable and suitable for decision support, they are represented by sequences of numbers as vectors. Every vector is called "**Domain Vector**" (**DV**). It is element of a nestable metric space which is called "**Domain Space**" (**DS**). The DS and every of its dimensions have a unique name (URL). A dimension can represent an unbranched value (usually a number) or again a DS to allow maximal reusage of existing definitions. DSs can be defined by all owners of web space. The URL of the DS definition also serves in the data (in every DV) as identifier of the DS. Every data element (DV) contains the URL of the complete standardized DS definition plus a sequence of values (vector). A two-dimensional example (Date with Bodyweight):

http://numericsearch.com/bw.xml; 2014-01-30; 83.914;

Abbreviation of URLs e.g. by local substitutes is possible. Here we do not anticipate the exact standard, but it becomes clear that an efficient and short syntax is possible.

3. Demonstration, last improvements and differences to other systems

A demo version of the search engine with test data is online [14]. Fig. 1 shows exemplary search data and Fig. 2 the result. The implementation again shows searchability of precise data in metric spaces. All DVs with selectable and freely defined quantitative data can be filtered from large data collections.

0	7.5		
1	2.5		
2			

Figure 1. Excerpt from the search mask of the implementation used for selection of records with similar data.

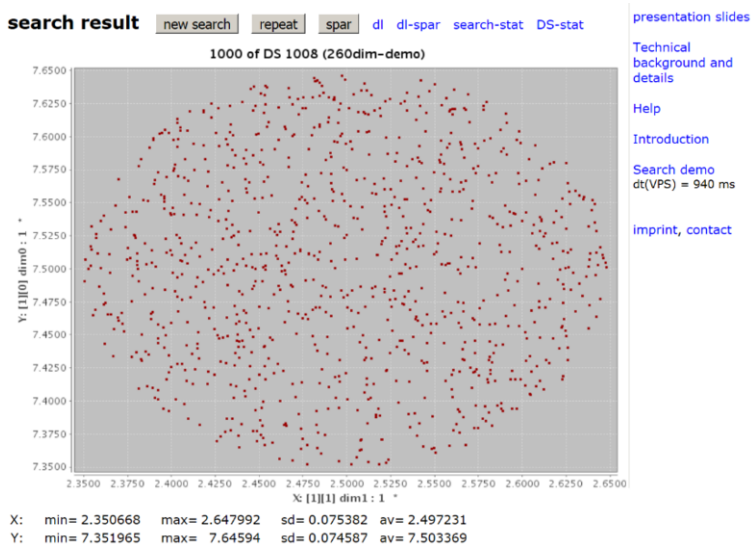

Figure 2. Excerpt of the search result. Shown are the 2 searched dimensions of the nearest 1000 records from a 260 dimensional Domain Space with 1500001 DVs with pseudorandom numbers between 0 and 10 in the 2 searched dimensions.

3.1. Most significant improvements in the last time

This can be seen when comparing v1 and v2 of [16]. The concept has been generalized. DVs can be used as uniform container of all definable information, see chapter 4.

3.2. Most significant difference to other systems

Most significant difference is the basal (language independent) numeric representation of information, combined with standardized (if wished multilingual) online definition of every dimension, possible everywhere on the web. If another systems introduces this and to puts emphasis on efficiency, a combination is possible.

4. Foundation

Usage of worldwide uniformly defined metric spaces like DSs is well justified and it seems appropriate to repeat the description of the foundation here:

Basically information (as well as decision) means selection within a given domain (value or definition set - on this basis e.g. in [10] the quantity of information is defined). Precondition of well defined information (or decision support) is that all speak and

think about the same domain. Its unambiguous global (online) definition is therefore recommendable for communication and interoperability. If we define an ordered domain, we can use numbers for identification of elements. If we also define a metric (distance function), we can quantify similarity and search in the domain. This leads to the proposal of Domain Spaces. A Domain Space (DS) is a user defined nestable metric space with unique name everywhere on the web (URL of definition) - for globally uniformly (language independent) defined searchable information (in DVs).

5. Application in medicine

DSs can cover all definable topics. Due to the focus of this paper here we describe usage of DSs in collections of medical original data for decision support. The proposed clinical procedure contains the following important steps:

1. The doctor makes a first principal diagnosis, e.g. ICD-10 [18]. SNOMED CT [2] would be also recommendable, if free. The code system should be reproducible, clear and free, because all connected data depend on it, and the value of the total data collection is growing during time without limit.
2. Using the code (later combined with additional finer quantitative findings) the software shows frequencies of fine (quantitative) diagnostics chosen by other doctors in such a case. This reduces the risk to overlook something important. The representation should be clear and the software interface user friendly.
3. The doctor decides about finer diagnostics. This includes collection of relevant collateral data e.g. about daily food intake, sports etc.
4. The multidimensional results of finer diagnostics are provided to the software (if possible, more and more automatically). The completeness, resolution and efficiency of such quantitative information it not nearly attainable by non quantitative code systems (example: GPS coordinates compared to postal addresses). Most important quantitative results are used as search criterion.
5. If enough patients with similar data can be found by the search engine, anonymously frequencies of further diagnoses with treatment decisions and associated results can be shown in this group of patients, like in a scientific study ("individual study").
6. Decision about further diagnostics or treatment is done and provided to software which can prepare, if wished, the draft of a medical report.
7. If necessary, later (with new data) continuation at 2. or even 1. If the patient (or parents) agree, the own codes, decisions and results remain pseudonymously in the data collection to enlarge it. The patient can delete the own data at any time in life. The mapping of patient name to patient number is deleted at the time of death. The unerased pseudonymous data remain in the data collection and become international property. All quantitative data (DVs) are elements of DSs which are online and internationally defined in standardized form to ensure interoperability. Definitions (of DSs) are referred by their URLs and periodically their backup is done. Depending on the patient's wishes, data are utilized statistically (anonymously). Data are not deleted because never "all" possibilities for medical evaluation are finished. So data collection and its medical value are growing more and more. The value of the data collection will become visible after delay.

6. Conclusion

Using DSs it is technically feasible to collect worldwide precise data from medical findings, therapeutic decisions and treatment results in a way that they are anonymously available at situations where similar decisions again have to be done. This can be used e.g. for creation of individual statistics from a group of patients with medical data similar to a given patient, to find the therapy which is associated in this individual group with the best results. New medical experience can be added to the worldwide database again and again.

References

[1] Abbasi, K., The missing data that cost $20 bn. BMJ, 348, g2695. http://www.bmj.com/content/348/bmj.g2695, 2014.
[2] Benson, T., Principles of health interoperability HL7 and SNOMED. Springer Science & Business Media, 2012.
[3] Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., ed. Feature Extraction, Foundations and Applications. Studies in Fuzziness and Soft Computing, Springer, Berlin, Heidelberg, 2006.
[4] Haas, P., Medizinische Informationssysteme und Elektronische Krankenakten. Springer, Berlin, Heidelberg, 2005.
[5] Hanauer, D. A., EMERSE: the electronic medical record search engine. In AMIA annual symposium proceedings. American Medical Informatics Association, (2006), p. 1189.
[6] HL7 Resource Observation - Examples, http://www.hl7.org/fhir/observation-examples.html, viewed January 2016.
[7] IBM Watson Health, http://www.ibm.com/smarterplanet/us/en/ibmwatson/health, viewed January 2016.
[8] Imler, T. D., Vreeman, D. J., & Kannry, J., Healthcare Data Standards and Exchange. In Clinical Informatics Study Guide. Springer International Publishing, (2016), 233-253.
[9] Jefferson, T., Jones, M., Doshi, P., & Del Mar, C., Neuraminidase inhibitors for preventing and treating influenza in healthy adults: systematic review and meta-analysis. Bmj, 339. http://www.bmj.com/content/339/bmj. ,(2009) b5106.
[10] Kolmogorov, A. N., Three approaches to the quantitative definition of information. Problems of information transmission, 1(1), (1968) 1-7.
[11] Kriegel, H.P., Kröger, P., Renz, M., Schubert, M., Metric spaces in data mining: applications to clustering. SIGSPATIAL Special Volume 2 Issue 2 (July 2010), http://dl.acm.org/citation.cfm?doid=1862413.1862423, (2010) 36-39.
[12] Orthuber, W., Dietze, S., Towards Standardized Vectorial Resource Descriptors on the Web. 2010. In Lecture Notes in Informatics, INFORMATIK 2010. Service Science - Neue Perspektiven für die Informatik. Band 2 P-176, http://subs.emis.de/LNI/Proceedings/Proceedings176/article6010.html, (2010) 453-458.
[13] Orthuber, W., Papavramidis E., Standardized Vectorial Representation of Medical Data in Patient Records, Medical and Care Compunetics 6, (2010) 153-166.
[14] Orthuber, W., Numeric Search - online implementation of vectorial description and search. (Online since July 2012) http://numericsearch.com, 2012.
[15] Orthuber, W., Exemplary DS definition for decision support in medicine (2014 without real data). http://numericsearch.com/w7s.jsp?i7=1012 , 2014.
[16] Orthuber, W., Uniform definition of comparable and searchable information on the web. arXiv preprint arXiv:1406.1065. http://arxiv.org/abs/1406.1065, 2016.
[17] Orthuber, W. How to make quantitative data on the web searchable and interoperable part of the common vocabulary. 2015 GI-Jahrestagung.
[18] Rubenstein, J. N., Painter, M. R., Painter, M., Schoor, R., & Baum, N. The 4 Questions to Ask and Answer Regarding ICD-10: Second of a 2-Part Series. Urology Practice, 2(2) , (2015) 65-68.
[19] Smits, M., Kramer, E., Harthoorn, M., & Cornet, R., A comparison of two Detailed Clinical Model representations: FHIR and CDA. EJBI, 11(2), 2015.
[20] W3C RDF Examples, http://www.w3schools.com/xml/xml_rdf.asp , viewed January 2016.
[21] Zezula, P., Amato, G., Dohnal, V., Batko, M. Similarity Search. The Metric Space Approach. Series: Advances in Database Systems, Vol. 32., Springer, Berlin, Heidelberg, 2005.