*eHealth Beyond the Horizon – Get IT There*
*S.K. Andersen et al. (Eds.)*
*IOS Press, 2008*



# Analysis of EHRs for research, quality management and health politics

Walter GALL,[a,1] Wilfried GROSSMANN,[b] Georg DUFTSCHMID,[a]
Thomas WRBA,[a] Wolfgang DORDA[a]
[a]*Section of Medical Information and Retrieval Systems,*
*Core Unit for Medical Statistics, and Informatics, Medical University of Vienna*
[b]*Institute for Scientific Computing, University of Vienna*

**Abstract.** Lifelong electronic health records can supply valuable information for research, quality management and health politics in addition to supporting treatment of patients. Based on experiences with scientific data analysis in a university hospital environment, requirements on cross-institutional analysis of electronic health records in a healthcare system are discussed. The concept of archetypes can play a key role in this context. Archetypes can be utilized in data analysis for visualization, semantic linkage and finally for standardized data transfer.

**Keywords.** Archival-repository systems for medical records-EPR-CPR-EMR, Standards, Data analysis-extraction tools, Epidemiological research Hospital IS

## Introduction

Introducing standardized electronic health records (EHR) is a strategic e-Health goal in Europe [1] and therefore in Austria as well [2]. Implementing this important project is going to influence *documentation, communication* and *analysis* of patient-related data. Services based on standardized *electronic health record architectures (EHRAs)* [3] are intended not only to improve treatment of individual patients but also to give a fresh impetus for research, quality management and health politics.

*Data analysis* may concern the EHRs of individual patients or may be conducted across patients. The former case is most likely to involve a physician or patient screening an EHR for relevant data in the context of medical treatment. The latter case will normally involve a variety of EHRs analyzed by an investigator or statistician in an effort to identify commonalities or differences.

The focus of the present article is to provide an analysis of requirements that a system for cross-patient and cross-institutional analysis should meet in a heterogeneous distributed healthcare system. Discussing all analysis requirements in detail is beyond the scope of this communication. Therefore we shall use what we consider to be the most important requirements for *local data analysis* in a *hospital environment* (as state

[1] Corresponding Author: Walter Gall, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria; E-mail: walter.gall@meduniwien.ac.at

of the art) as departure point and focus our analysis on the additional requirements needed for cross-institutional analysis within a *healthcare system environment*.

## 1. Methods

As a departure point, we take a look at data analyses in a hospital. Alongside scientific studies and requirement compilations such as [4] or [5], we mainly rely on the ArchiMed scientific retrieval system [6] for this purpose. This system allows heterogeneous medical data from different routine applications to be analyzed across patients together with data of clinical studies. The system has been used at the Austrian university clinics in Vienna and Graz for 10 years. Around 3000 data analyses are performed every year.

The goal of our work is to specify requirements for a retrieval system in the healthcare environment, taking into account the recommendations of the Austrian EHR feasibility study [7].

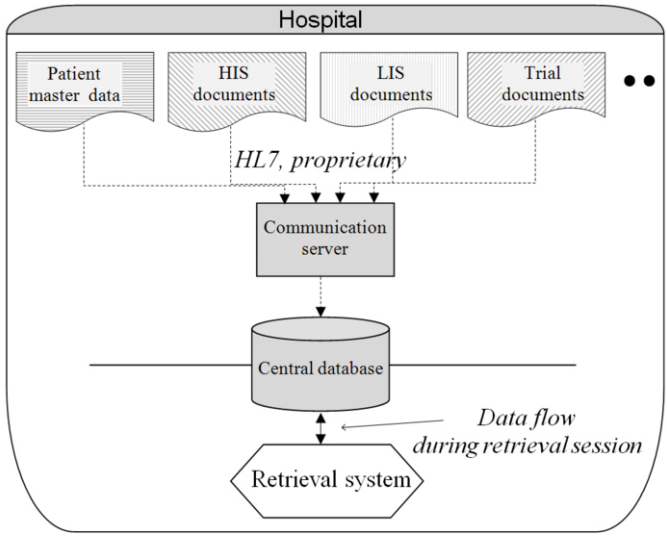


**Figure 1.** Data analysis in a hospital environment

Figure 1 illustrates a data retrieval system in a typical hospital environment. An ideal system should meet the following important *requirements:*

(1) **Data integration** by adopting data from documentation systems in a centralized data base that allows data to be analyzed based on a generic data model [8].
(2) **Support of the three typical retrieval steps** for clinical data, characterized by three sequential activities: *cohort formation* (searching for patients with

specific findings), *selection of variables* within the selected cohort (selecting the patient data of interest and processing them to form a statistical matrix) and *statistical analysis* (ranging from preparatory statistical functions up to utilizing the complete functionality of a statistical analysis package) [6]. The following requirements (3) and (4) are main features of the retrieval steps *cohort formation* and *selection of variables*.

(3) **Analysis of the time course** of a disease [9] by providing powerful temporal relations for selection conditions (e.g. "within 2 to 4 months after the surgery").
(4) **Text analysis** by using text operators for free text analysis and code systems such as ICD as thesaurus support. Additional algorithms are sometimes used to split up word compounds and to build textual roots [10] although the task of creating specialized textual knowledge bases has turned out to take an inordinate amount of time and effort.
(5) **Support of a large variety of user groups** by the bandwidth of the retrieval functionality, ranging from powerful components for flexible ad-hoc specifications of individual data analyses performed by power users down to prefabricated adaptable data analyses for novices.
(6) **Intuitive query formulation** by the use of a graphic query editor. This alternative to conventional line-based query formulation is routinely demanded but its implementation has usually been confined to prototypes.
(7) **Web-based analysis** through a Web interface to execute prefabricated data analyses. This functionality is not absolutely required within hospitals but is frequently used in the context of multicenter studies.
(8) **Data privacy** ensured by a comprehensive user and authorization system and anonymisation mechanisms for data export functions.

## 2. Results

Data retrieval performed in heterogeneous hospital environments that have developed over numerous years have a lot in common with data retrieval performed inside a nationwide healthcare system. However, the requirement profile cannot be directly transferred from one to the other. A comparison between Figure 1 and 2 shows that even the very mechanisms of data communication involved will differ greatly. In hospitals, data analyses are usually performed through a centralized database. In a nationwide EHR system, by contrast, documents need to be accessed that are spread over various places.

The following discussion focuses on the adaptations and extensions concerning the above-described requirements for a retrieval system in a nationwide EHR system environment.

(1) **Data integration.** Communication with the distributed repositories (assuming a structure such as IHE-XDS [11]) should be based on *archetypes* based on an *EHR information model* [12]. This will ensure *semantic interoperability*, which is a fundamental requirement for semantically correct data analysis.
(2) **Support of three typical analysis steps.** The extended information typology provided by the EHR has to be used. For *cohort formation* and *selection of*

*variables*, *organizational* (e.g. type of healthcare provider, federal state) and *context-related* (e.g. archetype) selection attributes should be available in addition to patient-related ones (e.g. patient, hospital stay). The requirements for statistical analysis remain unchanged.

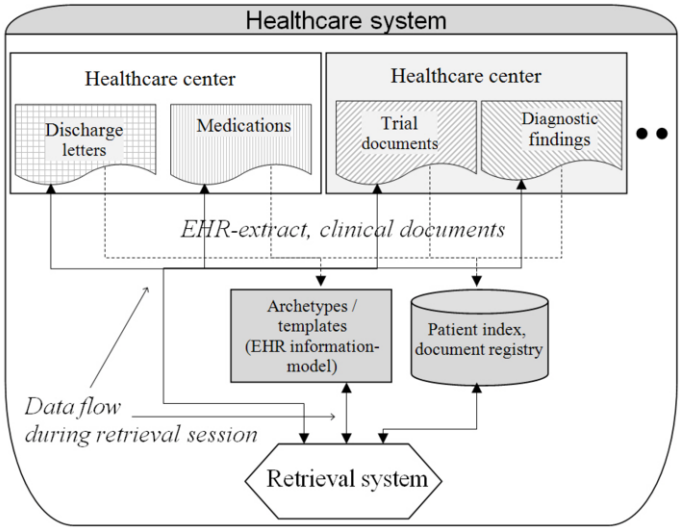


**Figure 2.** Data analysis in a countrywide EHR system environment

(3) **Analysis of the time course.** In the EHR the temporal logic has to be combined more closely with components pertaining to content (context-related archetypes). The question arises whether or not a specific temporal sequence can be evaluated jointly when the individual values were collected under different conditions, for example routine examination versus ergometry.

(4) **Text analysis.** Heterogeneity of data is increased due to the variety of data generators involved (e.g. physicians in private practice, university hospitals, laboratories). Analyses are not only focused on text parameters such as spelling or abbreviations but need to take into account the broader context of data compared to hospital environments. The use of *ontologies* [13] in conjunction with archetypes can contribute to supporting complex semantic analysis of medical terms.

(5) **Support of a variety of user groups.** The bandwidth of users does not become considerable larger in a healthcare system environment. However, the geographical extension is such that users can hardly be supported by conventional training and care. The on-line help functions need to cover all needs for assistance. As a result, they have to meet advanced quality requirements.

(6) **Intuitive query formulation.** Due to the larger complexity of cohort formation and selection of variables, the options of query formulation need to be expanded in proper way. Symbols and metaphors are gaining importance.

The EHR-based query languages that are currently being developed [14] need to be implemented in graphic format. Visualizing archetypes can both help to obtain a better overview of the context into which information is embedded and can serve to select variables.

(7) **Web-based data analysis.** Web applications are unavoidable in a nationwide EHR system environment. Technical challenges notably include the support of query formulations (e.g. by graphic means) based on Web technologies and performance requirements for accessing distributed data. In the communication domain, new possibilities arise with the introduction of asynchronous methods (e.g. Ajax).

(8) **Data privacy.** Whenever patient data leave a healthcare facility, data protection is a major concern. Investigators using such data inside a hospital frequently do so implicitly in a treatment context that will justify the analysis of data related to individuals. This context is usually lost once the data are accessed in a healthcare system environment. The data can only be analysed in an anonymized fashion. If identifiers are needed for complex multi-step queries (cohort formation, selection of variables), the requirements of *k-anonymity* must be met in addition to *anonymization* and *pseudonymization*. The requirements of *k-anonymity* are met if a cohort encompasses at least *k* data records with the same secondary identification characteristics [15]. When data groups are extracted, care must be taken to avoid any inadvertent inclusion of secondary identification characteristics that might not be consistent with the described criteria of k-anonymity.

## 3. Discussion

Introducing a standardized electronic health record is a defined strategic e-health goal in Europe. In the present article, a number of key requirements have been discussed that systems used for cross-patient analysis in a nationwide EHR system environment should meet. Numerous of these aspects will also apply to intra-patient data analysis as conducted by treating physicians or the patients themselves.

Naturally the requirements discussed in this paper can still be extended (e.g. to include result management), refined (e.g. anonymization algorithms) and examined for the time and effort that will likely go into their implementation. Numerous questions remain to be settled also with regard to *data privacy* as the most important aspect. Where should analysis steps be conducted? How can a document be opened to analyze the medical concepts used (archetypes) while ensuring anonymity at the same time?

Semantic integration of data is a quintessential task. *Archetypes* will play a key role in this connection. This concept is promoted by the three most important EHRA sources (CEN, HL7, openEHR). Their use builds a first step to render heterogeneous existing data interoperable [16] and hence analysable.

Selecting variables for analysis from complex data structures can be supported by visualizing archetypes with graphic user interfaces. Furthermore, *semantic links* can be used in conjunction with ontologies to extend a search through one archetype to other archetypes.

For example, if an analysis of data on diabetes mellitus according to patients is performed by searching for specific values of the medical history and of the blood sugar, the archetype “diabetes” could be utilized first to find the variables “medical

history" and "blood sugar". Furthermore, the search could be extended through the ontology section of the variable "blood sugar" to include an archetype "laboratory chemistry" containing the synonymous variable "blood glucose".

## 4. Conclusion

The aspect of data retrieval is frequently a stepchild in comparison with the aspect of data collection. It is not considered until data collection has already been completed. By that time, it is too late to take the requirements for data analysis into account. Even in standards for EHRAs, the "retrieval" aspect frequently falls into the category "future work". Retrieval-oriented requirements should already be considered when a standardized EHR system is introduced. Only then can this veritable gold mine of information be used efficiently for quality management, epidemiologic research and health politics.

## References

[1] Commission of the European Communities. e-Health - making healthcare better for European citizens: An action plan for a European e-Health Area. http://eur-lex.europa.eu/LexUriServ/site/en/com/2004/com2004_0356en01.pdf (accessed 2008-02-10).
[2] Austrian Federal Minister for Health and Woman. Excerpt from the 2005 Health Reform Act, Fed. Law Gaz. I no. 179/2004. http://www.bmgfj.gv.at/cms/site/attachments/3/4/1/CH0027/CMS1043931577060/health_care_quality_act.pdf (accessed 2008-02-10).
[3] Blobel B. Advanced EHR architectures--promises or reality. Methods Inf Med 2006;45(1):95-101.
[4] Safran C, Chute C. Exploration and exploitation of clinical databases. International Journal of Biomedical Computing, 1995; 39:151-156.
[5] Ammenwerth E, Buchauer A, Haux R. A requirements index for information processing in hospitals. Methods Inf Med. 2002; 41(4):282-8.
[6] Gall W, Sachs P, Duftschmid G, Dorda W. A retrieval system for the selection and statistical analysis of clinical data. Med Inform Internet Med. 1999; 24(3):201-12.
[7] Austrian Federal Minister for Health and Woman. Feasibility study ELGA. http://www.bmgfj.gv.at/cms/site/attachments/7/2/0/CH0513/CMS1169796766007/machbarkeitsstudie_elga.pdf (in German, accessed 2008-02-10).
[8] Dorda W, Wrba T, Duftschmid G, Sachs P, Gall W, Rehnelt C et al. ArchiMed: a medical information and retrieval system. Methods Inf Med. 1999; 38(1):16-24.
[9] Shahar Y, Combi C. Timing is everything. Time-oriented clinical information systems. West J Med. 1998; 168(2):105-13.
[10] Dorda W. Data-screening and retrieval of medical data by the system WAREL. Methods Inf Med. 1990; 29(1):3-11.
[11] IHE ITI Technical Committee. IHE IT Infrastructure Technical Framework, vol. 1 and vol. 2.
[12] Beale T. Archetypes and the EHR. Stud Health Technol Inform. 2003; 96:238-44.
[13] Smith B, Ceusters W. An ontology-based methodology for the migration of biomedical terminologies to electronic health records. AMIA Annu Symp Proc. 2005; 704-8.
[14] Ma C, Frankel H, Beale T, Heard S. EHR Query Language (EQL) – A Query Language for Archetype-Based Health Records. Stud Health Techn Inform. 2007; 129:397-401.
[15] Sweeney A. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002; 10(5): 557-570.
[16] Kalra D, Blobel G. Semantic Interoperability of EHR Systems. Stud Health Technol Inform. 2007; 127:231-45.