Integration of Genomic Data in Electronic Health Records

Opportunities and Dilemmas

U. Sax[1, 3], **S. Schmidt**[2, 3]

[1]Children's Hospital Informatics Program (CHIP), Harvard-MIT Division of Health Sciences and Technology, Children's Hospital Boston, Boston, Massachusetts, USA
[2]Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts, USA
[3]Harvard Medical School, Boston, Massachusetts, USA

Summary

Objectives: In this paper we give an overview about the challenge the postgenomic era poses on biomedical informaticists. The occurrence of new (genomic) data types necessitates new data models, new viewing metaphors and methods to deal with the disclosure of genomic data. We discuss integration issues when inferring phenotype and genotype data. Another challenge is to find the right phenotype to genotype data in order to get appropriate case numbers for sound clinical genotype-phenotype inference studies.
Methods: Genomic data could be integrated in an Electronic Health Record (EHR) in several ways. We describe patient-centered and pointer-based integration strategies and the corresponding data types and data models. The inference mechanisms for the interpretation of raw data contain different agents. We describe vertical, horizontal and temporal agents.
Results: We have to deal with several new data types, not being standardized for EHR integration. Genomic data tends to be more structured than phenotype data. Beyond the development of new data models, vertical, horizontal and temporal agents have to be developed in order to link genotype and phenotype. As the genomic EHR will contain very sensitive data, confidentiality and privacy concerns have to be addressed.
Conclusions: Given the necessity to capture both environment and genomic state of a patient and their interaction, clinical information systems have to be redesigned. While genotyping seems to be automatable easily, this is not the case for clinical information. More integration work on terminologies and ontologies has to be done.

Keywords

Human genome, phenotype, genotype, personal health record, hospital information systems, medical informatics, bioinformatics

Methods Inf Med 2005; 44: 546–50

Introduction

While the human genome has been sequenced [1, 2], the challenging part begins now, as we still do not know the function of many genomic regions. The post-genomic era of medicine not only challenges patients, clinicians and the healthcare system, but also will have a huge impact on healthcare information systems.

Genomic patient information such as the results of genotyping disease-associated genes will likely be part of the patient's healthcare record soon, while only few clinical information systems are prepared to deal with genomic data [3].

Historically, clinical patient data, research data as well as clinical study data were stored in different databases and systems. This separation of clinical data and research data constrains new insight more than ever before. In order to find the relationship between genotype and phenotype, we need clinical data as well as research data accessible for inference mechanisms and clinical studies.

As the genotype does not solely capture the individual patient state, we additionally need to assess and quantify environmental influences. This comprises the patient history, physical condition, laboratory studies and imaging data [4, 5].

Therefore, we need new data models and data structures to cope with the integration of genomic data in the Electronic Health Record (EHR) as well as inference mechanisms to connect genotype and phenotype data.

Electronic patient records have a long history in medical informatics [6]. Mainly dealing with billing data in the early years, a current EHR supports many kinds of phenotypic patient data [7].

Currently, genomic data is not represented in the EHR standard models, but a HL7 Special Interest Group (SIG) is creating a HL7 Clinical Genomics Model [8]. We have to deal with an inconvenient type of data, because predictions and interpretations drawn from a patient's DNA sequence will have to be repeated frequently as research gains new insights. Therefore the raw data should be easily accessible.

Finally genomic data is considered as highly confidential and has to be protected from unauthorized access [9, 10]. Early genomic databases like in Iceland or Estonia raise severe ethical concerns [11].

Methods

In the first step we had a close look at data types and data formats in biomedical informatics. We then examined the most relevant standard formats for phenotype and genotype data. Finally we assessed the available data models for the ability to represent both genomic and phenotypic data.

Data Types and Data Formats

Genomic data are mostly presentable in eXtensible Markup Language (XML) [12]. Nucleotide and protein sequences are simple ASCII streams. GenBank wraps these data in a descriptive XML header [13]. Protein data (amino acid sequences) might be derived from genomic data directly (annotation of the genome). Uniprot (formerly SwissProt) delivers a popular format to store protein data [14] whereas the Protein DataBase (PDB) defines a format to describe protein structure [15]. In case of microarray data, it can be delivered in the MAGE-ML format [16] corresponding to the "Minimum Information about Microarray Experiments" (MIAME) recommendation. Falling prices of the tests for Single Nucleotide Polymorphisms (SNPs) raises their popularity [17]. dbSNP uses a widely accepted XML-format for storing SNPs [18].

Biomedical databases-designers are confronted with moving targets, as they struggle from the vast amounts of data being submitted including updates and redundancy [19].

Most of the aforementioned data are linked in so called meta-databases. GeneCards [20], the genetic association database [21] and LokusLink (will soon be transformed in "Entrez Gene"), offer curated and non-redundant, highly linked data. In general, the NCBI databases – as well as their European pendants [22] – are highly structured and the primary keys linked to the corresponding databases [23].

Data Models

The Health Level Seven (HL7) Clinical Genomics Special Interest Group (cgSIG) proposed extensions to the HL7 Reference Information Model (RIM), as well as for the clinical study-centered Clinical Data Interchange Standards Consortium (CDISC) standard [8]. The HL7 cgSIG proposes a new RIM-based genotype model, which will be implemented as reusable common message element type (CMET).

A patient's genotype consisting of individual alleles is hooked up at the top layer of the model with the possibility of versioning it. The model allows storing the allele sequence as well as associated observations like individual Single Nucleotide Polymorphisms (SNPs), mutations, and gene expression data. The alleles can furthermore be associated with haplotypes and polymorphisms. The strength of the HL7 proposal is the possibility to associate the alleles with clinical phenotype observations like diseases, risk factors and adverse drug events.

Closely related to the HL7 RIM biomedical data can be represented in Clinical Document Architecture (CDA) documents.

Results

Integration of genomic data in an EHR seems to be more difficult than for example the integration of clinical chemistry lab data. We do not only need the results, but also the raw data, because the data has to be reanalyzed frequently, as new insight could lead to a different result. In the following we focus on a strategy for the integration of genomic data, their representation and on inference mechanisms towards the synthesis of novel insights.

Integration Aspects

For a patient-centered data integration, we need the lab meta-data such as the specimen, used methods and corresponding interpretations. It seems to be crucial to include the unprocessed data from sequencing labs or the results of microarray experiments on patients as well.

One integration strategy could be to include genomic meta-data directly in the EHR and to provide a pointer on the raw data due to the huge amount of data. But this could endanger the longitudinal characteristic of the record, as many of the current genomic databases only have a short half-life, especially concerning data formats.

As the above-mentioned data formats allow wrapping the raw data and annotating them with meta data, a good strategy could be to include the raw data as well as the pointer to the external resources.

Representation of Genomic Data

The HL7 clinical genomics SIG proposes an individual genotype model; genomic data can be represented in an EHR as a new observation class [24] in the data model.

On the communication level, a clinical genomics report could be represented as a self-contained document according to the HL7 Clinical Document Architecture (CDA). CDA Revision 1 was approved as an ANSI standard in 2000 [25]; Revision 2 was ANSI-approved in May 2005 [26]. CDA defines three levels; in level one only the document header is fixed with a certain set of document and object identifiers, whereas the document body may contain any XML data choosing the unstructured body option. Level 2 documents additionally contain a structured body with certain sections, which have to be derived from the HL7 Reference Information Model (RIM) in Level 3 Documents.

In the CDA header the document type as well as the universal observation identifiers have to be defined according to the Logical Observation Identifier Names and Codes (LOINC) document types [27]. Furthermore the data format and the current version have to be registered and to be noted in the document header in order to be able to present the document accordingly. European projects like PICNIC [28] and SCIPHOX [29] successfully make use of CDA level one documents. This practice would have the advantage to be able to represent any genomic data format without the necessity to reinvent a new document structure. As the standardization process on genomic issues proceeds, the documents could be refined as level three documents successively.

Integration Needs a Common Ontology

By integration of genotype and phenotype data we address the influx of either genomic data on clinical data, prognosis, and therapy or vice versa. Likely there will be two new directions of clinical studies in the future. The one direction would lead from phenotype to genotype what means, that patients with particular clinical traits are screened for genes known or suspected to be involved in that disease to identify variations. The other direction would lead from genotype to phenotype, suggesting a screening of large numbers of individuals at particular genetic loci and then looking for phenotypic communalities or differences.

This approach is well known in industry, as industry-IT has to deal with vast data amounts from various data sources as well. Data mining tools ease the selection of interesting data in data warehouses and deliver the appropriate slices and dices of data for further examination. For example IBM tries to apply this technique to healthcare to allow some new insight [30].

However, both directions of post-genomic clinical studies and the application of data mining methods imply standardized patient data and standardized terminology both on clinical and on genomic side being accessible for knowledge-finding tools. This could actually be the most difficult part of the integration effort.

Though most of the described genomic data can be accessed as XML data, the content is partially not well structured for the purpose of automatic inference with clinical data. Gene Expression Omnibus (GEO) [23] for example holds an interesting set of Microarray experiment data, but the description of used material, kind of treatment in the affected group, and the affiliation of the data set to treatment or control is not standardized. All these features are combined in the free text field "description". The Microarray Gene Expression – Markup Language (MAGE-ML) [16] defines all the relevant fields like "BioMaterial" and "ExperimentalDesign" – but they are not in use.

On the clinical side, an even bigger obstacle is the lack of ontological compatibility, as the lion's share of phenotype data is not encoded for automated processing. While the data is still partially captured in paper files, electronic patient data is scattered over several information systems in a variety of different data formats. Ideally clinical data could be encoded using terminologies like LOINC [27], classifications like the International Classification of Diseases (ICD), and the appropriate mapping to the Unified Medical Language System (UMLS) [31].

So the first step to approach the problem is to work on terminology and ontology and transform the content in a uniformed representation like UMLS. First attempts show the complexity of this task [3, 8, 32-34]. The second step would be to build an infrastructure to capture all phenotype data of a patient in a place like a Personal Health Record (PHR) [35].

Agents

The current interpretation of the raw data for one patient has to be versioned, as with new insights in science these interpretations could change over time. We need electronic agents, similar to the ideas in early Personal Health Record papers [36] and described in some bioinformatics integration approaches [37]. These agents frequently analyze the given data using recent scientific knowledge, to keep the data sets and their interpretation up to date.

Vertical agents frequently analyze the data of one patient; horizontal inference agents infer genotype and phenotype information using EHR data from many patients in order to generate new knowledge for genomic epidemiology. Finally temporal agents analyze patient data to create phenotype time series that can be related to a genotype.

Privacy

As the enriched EHR will contain very sensitive genomic data, confidentiality and privacy concerns have to be addressed for two reasons: the genomic data is much more predictive of the patient's health status than any other test, and the genome is uniquely identifiable [5, 9-11]. Possibly it is sufficient to store only significant parts of the sequences to address the privacy issues. The difficult part will be to find the trade-off between privacy and disclosure [38]. Additionally to the explicit patient consent to store and to examine her or his genomic data, the data has to be specially protected.

Discussion

The representation of genomic data in EHRs seems to be feasible with HL7 CDA in general. We implemented a CDA document containing the result of a SNP test in HL7 CDA Revision 1 Level 1 (R1 L1).

As CDA is on the leap from Revision 1 to Revision 2 [26], there remains some uncertainty in the sustainability of R1 documents. Furthermore, the results of the Clinical Genomics Special Interest Group [8] (CG SIG) are not yet implemented in the HL7 RIM and therefore not in the corresponding CDA XML schemas.

We experienced good support in the area of billing or insurance-related document types – i.e. phenotype data. The area of handling genomic data is still evolving. For example there are no LOINC codes and Object Identifiers available yet [39] to code genomics-related section captions.

From the viewpoint of semantic interoperability, UMLS-mapping of expert terminology is very useful. As any CDA document element should be specified with an OID and a LOINC code and LOINC is represented in the UMLS Metathesaurus [31, 40], a concept mapping within different CDA documents seems possible at least for phenotype data at this time. Some more work has to be done in order to get sustainable genomic CDA documents.

Conclusion

The opportunities of the combining genomic data with phenotype data are obvious, as they allow a new type of clinical investigations in order to get new insights concerning diagnoses, prognosis and therapeutics. These studies are a crucial part of the National Institutes of Health (NIH) roadmap, meanwhile four National Centers for Biomedical Computing got funded in 2004 [41, 42].

The dilemma in 2005's Biomedical Informatics is that the data is present, but the enormous amounts of data are not easily linkable with each other. Systems to capture genotype data are thoroughly divided from systems for clinical phenotype data capture. Unfortunately the awareness of the necessity for changes in healthcare IT-systems is hardly to be found yet. For example a recent issue of BMJ addressed the future of healthcare informatics – not one article covered the impact of the post-genomic era [43].

Given the necessity to capture both environment and genomic state of a patient and their interaction, clinical information systems will have to be redesigned. Early examples of these systems can be seen in IBM's Genomics Messaging System [44], being rolled out in the Mayo clinic recently [30].

An even bigger obstacle than database integration is the lack of terminological and ontological compatibility, which could be solved by means of a uniformed representation like the Unified Medical Language System (UMLS) [31].

Beyond the development of new data models and ontological challenges, vertical (patient-centered), horizontal (study-centered) and temporal (time series) agents have to be developed in order to link genotype and phenotype. While genotyping seems to be automatable easily, this is not the case for clinical information. These agents allow applying new insights in genomic medicine to the present genomic data.

Finally, ethical and privacy concerns have to be addressed. Patients as well as health providers need education on this new methodology. Security-related procedures like patient consent as well as access control need to be established.

Acknowledgments

This work was supported by Deutsche Forschungsgemeinschaft (DFG, SA1009/1-1) and by the National Institutes of Health through contract N01-LM-3-3515 from the National Library of Medicine.

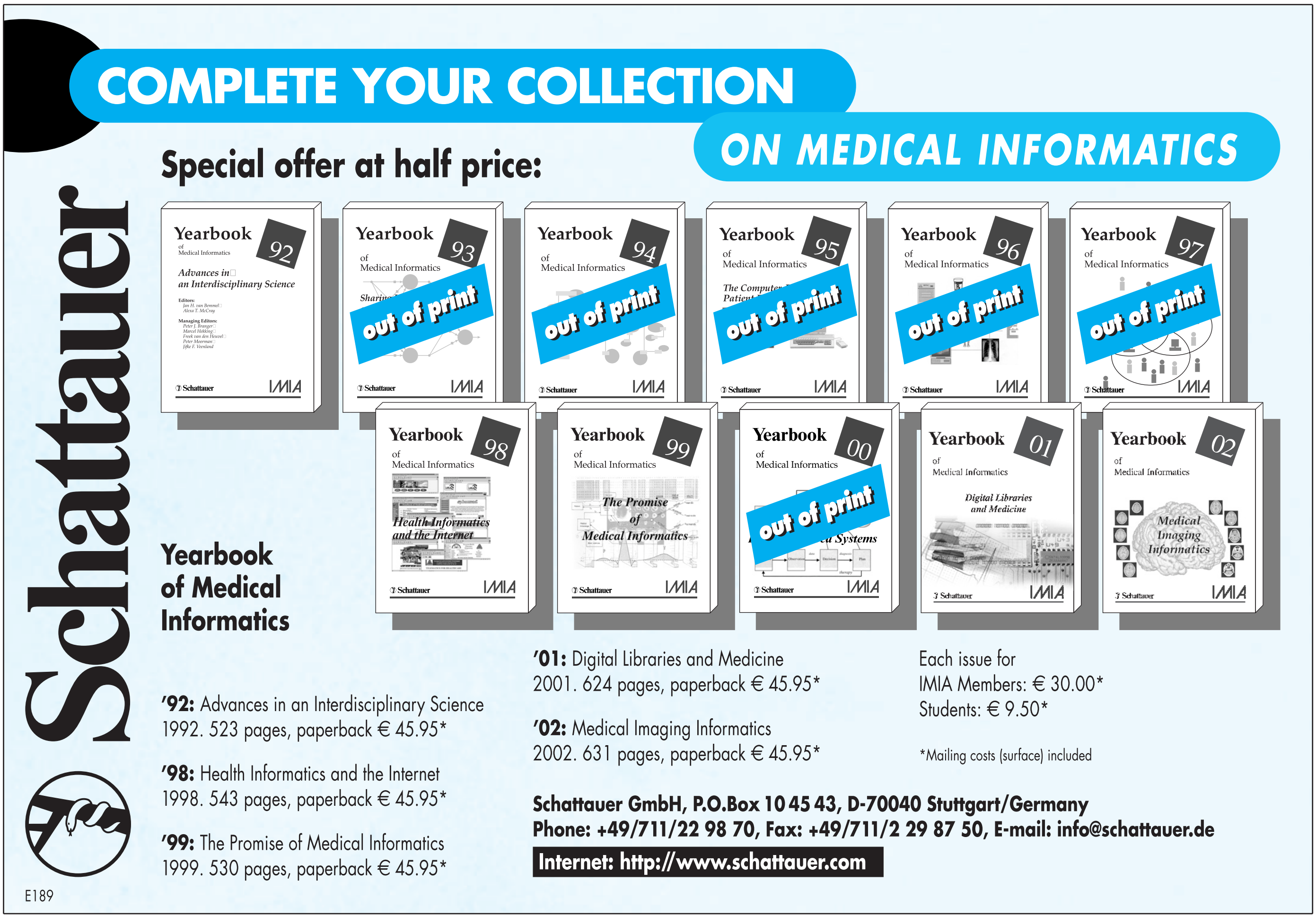
We would like to thank Zak Kohane for his critical review and some very interesting discussions as well as Martina Henn-Sax for her critical discussion and her patience.

References

1. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001; 291 (5507): 1304-51.
2. Istrail S, Sutton GG, Florea L, et al. Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci USA 2004; 101 (7): 1916-21.
3. Cerner Expands Industry Lead with Genome-Enabled Information System. Cerner Solutions. Available at: http://www.cerner.com/public/NewsReleases_1a.asp?id=257&cid=220. Accessed February 10, 2005.
4. Ford JH, 2nd, Turner A, Yoshii A. Information requirements of genomics researchers from the patient clinical record. J Healthc Inf Manag 2002; 16 (4): 56-61.
5. Kohane IS. Bioinformatics and clinical informatics: the imperative to collaborate. J Am Med Inform Assoc 2000; 7 (5): 512-6.
6. Giere W. Electronic Patient Information – Pioneers and MuchMore; A Vision, Lessons Learned, and Challenges. Methods Inf Med 2004; 43 (5): 543-52.
7. Klar R. Selected Impressions on the Beginning of the Electronic Medical Record and Patient Information. Methods Inf Med 2004; 43 (5): 537-42.
8. HL7 Clinical Genomics SIG. HL7 Clinical Genomics SIG San Diego Meeting Minutes Jan 21-22, 2004. HL7. Available at: http://tinyurl.com/4ducz. Accessed February 10, 2005.
9. Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J Biomed Inform 2004; 37 (3): 179-92.
10. Malin BA. An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future. J Am Med Inform Assoc Nov 18, 2004.
11. Kaiser J. Biobanks. Population databases boom, from Iceland to the U.S. Science 2002; 298 (5596): 1158-61.
12. W3C WWWC. Extensible Markup Language (XML). Available at: http://www.w3.org/XML/, 2003.
13. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2003; 31 (1): 23-7.
14. Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003; 31 (1): 365-70.
15. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. J Mol Biol 2003; 334 (4): 793-802.
16. Spellman PT, Miller M, Stewart J, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 2002; 3 (9): RESEARCH0046.
17. DNA Sequencing Costs Continue to Decline. FuturePundit. Available at: http://www.futurepundit.com/archives/002038.html. Accessed 14.01.2005.
18. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001; 29 (1): 308-11.
19. Buckingham S. Bioinformatics: data's future shock. Nature 2004; 428 (6984): 774-7.
20. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 1998; 14 (8): 656-64.
21. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet 2004; 36 (5): 431-2.
22. Bioinformatics at EMBL Heidelberg. European Molecular Biology Laboratory. Available at: http://www-db.embl.de/jss/EmblGroupsHD/serv_0?t=1&p=1#serv34. Accessed February 10, 2005.
23. Wheeler DL, Church DM, Edgar R, et al. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res 2004; 32 Database issue: D35-40.
24. Shabo A. Reusable Genotype R-MIM V0.4. HL7 Clinical-Genomics SIG. 2004-03-14. Available at: http://www.hl7.org/library/committees/clingenomics/HL7-Clinical-Genomics-Genotype-Model-0.4.zip. Accessed February 10, 2005.
25. Dolin RH, Alschuler L, Beebe C, et al. The HL7 Clinical Document Architecture. J Am Med Inform Assoc 2001; 8 (6): 552-69.
26. Dolin RH, Altschuler L, Boyer S, Beebe C, Behlen FM, Biron PV. HL7 Clinical Document Architecture (Release 2.0). Health Level Seven, Inc. Available at: http://hl7.org/library/Committees/structure/CDA.ReleaseTwo.CommitteeBallot03.Aug.2004.zip. Accessed February 9, 2005.
27. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003; 49 (4): 624-33.
28. Bray B. The PICNIC approach to regional care networks. Stud Health Technol Inform 2003; 96: 80-7.
29. Heitmann KU, Schweiger R, Dudeck J. Discharge and referral data exchange using global standards – the SCIPHOX project in Germany. Int J Med Inf 2003; 70 (2-3): 195-203.
30. Snow D. Mayo Amassed Mounds of Data. Wired News. Available at: http://www.wired.com/news/medtech/0,1286,61633,00.html?tw=wn_tophead_4. Accessed February 10, 2005.
31. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004; 32 Database issue: D267-70.
32. Holloway E. Meeting Review: From Genotype to Phenotype: Linking Bioinformatics and Medical Informatics Ontologies. Comp Funct Genom 2002; 2002 (3): 447-50.
33. Verschelde J-L, Dos Santos MC, Deray T, Smith B, Ceusters W. Ontology-assisted database integration to support natural language processing and biomedical data-mining. Journal of Integrative Bioinformatics, 15.01.2004 2004(0001, 2004).
34. Klein TE, Altman RB. PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. Pharmacogenomics J 2004; 4 (1): 1.
35. Riva A, Mandl KD, Oh DH, et al. The personal internetworked notary and guardian. Int J Med Inf 2001; 62 (1): 27-40.
36. Guardian Angel – Personal Lifelong Active Medical Assistant. MIT CDM. Available at: http://www.ga.org/ga/, February 10, 2005.
37. Karasavvas KA, Baldock R, Burger A. Bioinformatics integration and agent technology. J Biomed Inform 2004; 37 (3): 205-19.

38. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. Science 2004; 305 (5681): 183.
39. HL7 OID Registry. HL7 Inc. Available at: http://www.hl7.org/oid/mem_index.cfm. Accessed February 10, 2005.
40. FAct Sheet UMLS Metathesaurus. National Library of Medicine. Available at: http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html. Accessed February 10, 2005.
41. Biomedical Information Science and Technology Initiative (BISTI): National Centers for Biomedical Computing. National Institutes of Health. Available at: http://www.bisti.nih.gov/ncbc/index.cfm. Accessed February 10, 2005.
42. Kohane I, Glaser J. Informatics for Integrating Biology and the Bedside (I2B2). Available at: http://www.i2b2.org/index2.html. Accessed February 10, 2005.
43. Jadad AR, Delamothe T. What next for electronic communication and health care? Bmj 2004; 328 (7449): 1143-4.
44. IBM_Research. Genomics Messaging System (GMS). IBM Haifa Labs. Available at: http://www.haifa.il.ibm.com/projects/software/imr/gms.html. Accessed February 10, 2005.

Correspondence to:
Ulrich Sax, PhD, Assistant Professor
Department of Medical Informatics
CIOffice Medical Research Networks
Robert-Koch-Straße 40
37075 Göttingen
Germany
E-mail: usax@med.uni-goettingen.de


Schattauer
Yearbook of Medical Informatics 92
Advances in an Interdisciplinary Science
Yearbook of Medical Informatics 93
Yearbook of Medical Informatics 94
Yearbook of Medical Informatics 95
Yearbook of Medical Informatics 96
Yearbook of Medical Informatics 97
Yearbook of Medical Informatics 98
Health Informatics and the Internet
Yearbook of Medical Informatics 99
The Promise of Medical Informatics
Yearbook of Medical Informatics 00
Yearbook of Medical Informatics 01
Digital Libraries and Medicine
Yearbook of Medical Informatics 02
Medical Imaging Informatics
out of print
Schattauer
IMIA