Journal of Biomedical Informatics 45 (2012) 82–92

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Incorporating personalized gene sequence variants, molecular genetics knowledge, and health knowledge into an EHR prototype based on the Continuity of Care Record standard

Xia Jing [a,*], Stephen Kay [b], Thomas Marley [c], Nicholas R. Hardiker [d], James J. Cimino [a]

[a] Laboratory for Informatics Development, NIH Clinical Center and National Library of Medicine, Bethesda, MD, USA
[b] School of Health, Sport & Rehabilitation Sciences, University of Salford, Salford, Greater Manchester, UK
[c] Independent Consultant, Manchester, UK
[d] School of Nursing & Midwifery, University of Salford, Salford, Greater Manchester, UK

ARTICLE INFO

Article history:
Received 27 April 2011
Accepted 4 September 2011
Available online 17 September 2011

Keywords:
Molecular Genetic Information
Sequence variants
Electronic health record
Personalized information
Standards
Information filters

ABSTRACT

Objectives: The current volume and complexity of genetic tests, and the molecular genetics knowledge and health knowledge related to interpretation of the results of those tests, are rapidly outstripping the ability of individual clinicians to recall, understand and convey to their patients information relevant to their care. The tailoring of molecular genetics knowledge and health knowledge in clinical settings is important both for the provision of personalized medicine and to reduce clinician information overload. In this paper we describe the incorporation, customization and demonstration of molecular genetic data (mainly sequence variants), molecular genetics knowledge and health knowledge into a standards-based electronic health record (EHR) prototype developed specifically for this study.

Methods: We extended the CCR (Continuity of Care Record), an existing EHR standard for representing clinical data, to include molecular genetic data. An EHR prototype was built based on the extended CCR and designed to display relevant molecular genetics knowledge and health knowledge from an existing knowledge base for cystic fibrosis (OntoKBCF). We reconstructed test records from published case reports and represented them in the CCR schema. We then used the EHR to dynamically filter molecular genetics knowledge and health knowledge from OntoKBCF using molecular genetic data and clinical data from the test cases.

Results: The molecular genetic data were successfully incorporated in the CCR by creating a category of laboratory results called "Molecular Genetics" and specifying a particular class of test ("Gene Mutation Test") in this category. Unlike other laboratory tests reported in the CCR, results of tests in this class required additional attributes ("Molecular Structure" and "Molecular Position") to support interpretation by clinicians. These results, along with clinical data (age, sex, ethnicity, diagnostic procedures, and therapies) were used by the EHR to filter and present molecular genetics knowledge and health knowledge from OntoKBCF.

Conclusions: This research shows a feasible model for delivering patient sequence variants and presenting tailored molecular genetics knowledge and health knowledge via a standards-based EHR system prototype. EHR standards can be extended to include the necessary patient data (as we have demonstrated in the case of the CCR), while knowledge can be obtained from external knowledge bases that are created and maintained independently from the EHR. This approach can form the basis for a personalized medicine framework, a more comprehensive standards-based EHR system and a potential platform for advancing translational research by both disseminating results and providing opportunities for new insights into phenotype-genotype relationships.

1. Introduction

Personalized medicine has been defined as "a form of medicine that uses information about a person's genes, proteins, and environment to prevent, diagnose, and treat disease" [1]. The potential clinical significance of a patient's particular mutations and genes, which we refer to as *genetic data*,[1] demands that they be integrated into the clinical environment in a meaningful way.

* Corresponding author. Address: Building 10 (CRC, Room 5-2740), 10 Center Drive, NIH Clinical Center, Bethesda, MD 20892, USA.
E-mail address: xia.jing@nih.gov (X. Jing).

[1] For definitions of this term and other technical terms used throughout the paper, see the Glossary.

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.09.001

However, the characteristics of genetic data and the specific characteristics of the clinical contexts (such as the normally limited physician–patient interview and consultation time), create major challenges to such integration.

Genetic data is a new type of laboratory data that is revealed by testing a subject's genetic characteristics at the molecular level, such as gene sequence variations that cause diseases. Consider, for example, the genetic variations that occur in patients with cystic fibrosis. The *CFTR* (cystic fibrosis transmembrane conductance regulator) gene is responsible for coding the sequence of the chloride channel-CFTR protein. CFTR protein is composed of a chain of 1480 amino acids. Among the 1600 reported mutations of *CFTR*, one particular mutation, CFTR G551D, results in the replacement of glycine (G) with aspartic acid (D) at position 551 of the amino acid chain.

Genetic sequencing can identify this mutation in a patient's genome and thus the presence of CFTR G551D can be included in the patient's health record. However, this mutation represents only one data point among the 3 billion base-pairs in the patient's entire genome. As whole genome sequencing becomes practical and affordable, clinicians and patients are likely to have many questions when such data are routinely included in the health record.

The integration of genetic data in decision making or risk assessment is critical for clinical care; however, this must be accomplished without overwhelming clinicians and patients with all available information. Simply including the entire genome in the health record is unlikely to lead to successful use of the data. Seeing that a patient has CFTR G551D tells clinicians nothing about the health implications of this mutation for the patient, for example whether the patient is more prone to pneumonia or pancreatitis; relying on the clinician's memory for each of the 1600 variants is unlikely. Proper integration of the genome into the health record requires judicious selection of relevant portions for inclusion or highlighting, as well as incorporation of, or access to, available *genetics knowledge* and *health knowledge* required for interpretation of the data.

Integration of genetic data into the electronic health record (EHR) is both a challenge and an important research topic [2–4]. The provision of personalized information at the point of care is one approach to tackling information overload in the process of personalized care [5]. Some reviews and perspectives [3,6,7] discuss the challenges; however, original investigations into integrating genetic data into the EHR environment are lacking.

In this research we explore how a patient's genetic data (mainly sequence variants), as well as general genetics knowledge and health knowledge about specific genetic findings can be integrated into a standards-based EHR data model and how genetics knowledge and health knowledge can be customized and displayed via an EHR prototype. The informatics problems we address in this research are (1) whether it is feasible to incorporate genetic data into a standard EHR data model in computable form and (2) how to filter genetics knowledge and health knowledge in a dynamic, patient-specific manner within an EHR. The major focus of this paper is the implementation of the steps needed to construct an EHR prototype that demonstrates this approach, including the personalization and presentation of genetics knowledge and health knowledge via the prototype.

2. Background

In this section, we describe the components that we bring together to construct the EHR prototype that embodies genetic data, genetics knowledge, and health knowledge.

2.1. OntoKBCF

Previous papers have introduced the construction of OntoKBCF [8,9], an ontological knowledge base for cystic fibrosis (CF). The major content in OntoKBCF includes gene therapy with regard to CF, time-related CF descriptions, the CF-related Cochrane review conclusions, the most common CFTR mutations, and the characteristics of those mutations. OntoKBCF provides relationships between genetics knowledge with health knowledge regarding CF, including semantic meanings of concepts and clinical interpretations of test results. The types of sequence variants in OntoKBCF include: substitution, deletion and insertion for both DNA and proteins.

The knowledge facts in OntoKBCF have been organized according to *time* and *problems* using the Web Ontology Language (OWL) [10]. Time facts relate to how CF may present at different ages of the subject, while problem facts are used to represent most common CFTR mutations, the Cochrane review conclusions and questions about gene therapy. The granularity of OntoKBCF ranges from atomic concepts (such as coughing) to the complex knowledge facts (such as the typical descriptions for the adolescent female CF patient).

For example, the semantic meaning of the aforementioned CFTR G551D variant includes an amino acid change, a mutation position and mutation results (Fig. 1). The clinical significance for patients with CFTR G551D includes knowledge facts such as the finding that CFTR G551D has a higher prevalence in English(ethnicity) CF patients [11] and that such patients are more prone to develop pancreatic insufficiency (see Fig. 2). This knowledge in OntoKBCF serves as a resource of genetics knowledge and health knowledge for the EHR prototype.

2.2. Genetic data

Genetic data differ from traditional laboratory data in several important ways:

(1) Complexity:
   (a) the precise mechanisms by which genetic data correspond to disease status are usually not fully understood and characterized;
   (b) the volume and diversity of genetic data are increasing rapidly. Although genetic data are being recorded with more detailed metadata and with better structure, the semantic meaning of the strings (i.e. the expression form of the genetic data) – and thus the clinical interpretation of the strings - are not easily understood; recent recommendations about reporting CFTR-related disorders requires that the interpretation of the genetic data (including the clinical significance of the detected mutations) be part of the standard report [13]. Therefore, including the interpretation of the genetic data is also a requirement according to domain experts.
   (c) gold standards (such as laboratory reference values) for interpreting, reviewing and comparing genetic data are lacking.

(2) Implications for personalized medicine: genetic data have huge potential for more specific individualized diagnosis, treatment and prognosis. Although the clinical mechanisms related to genetic data may not be well understood, the clinical significance of genetic data may nevertheless be recognized. For example, the *EGFR* (epidermal growth factor receptor) mutation is a useful biomarker and selection criteria in the therapy of non-small cell lung cancers and the test for this mutation has been recommended for routine practice [14] even though the precise relationship of the mutation and the disease are not known.

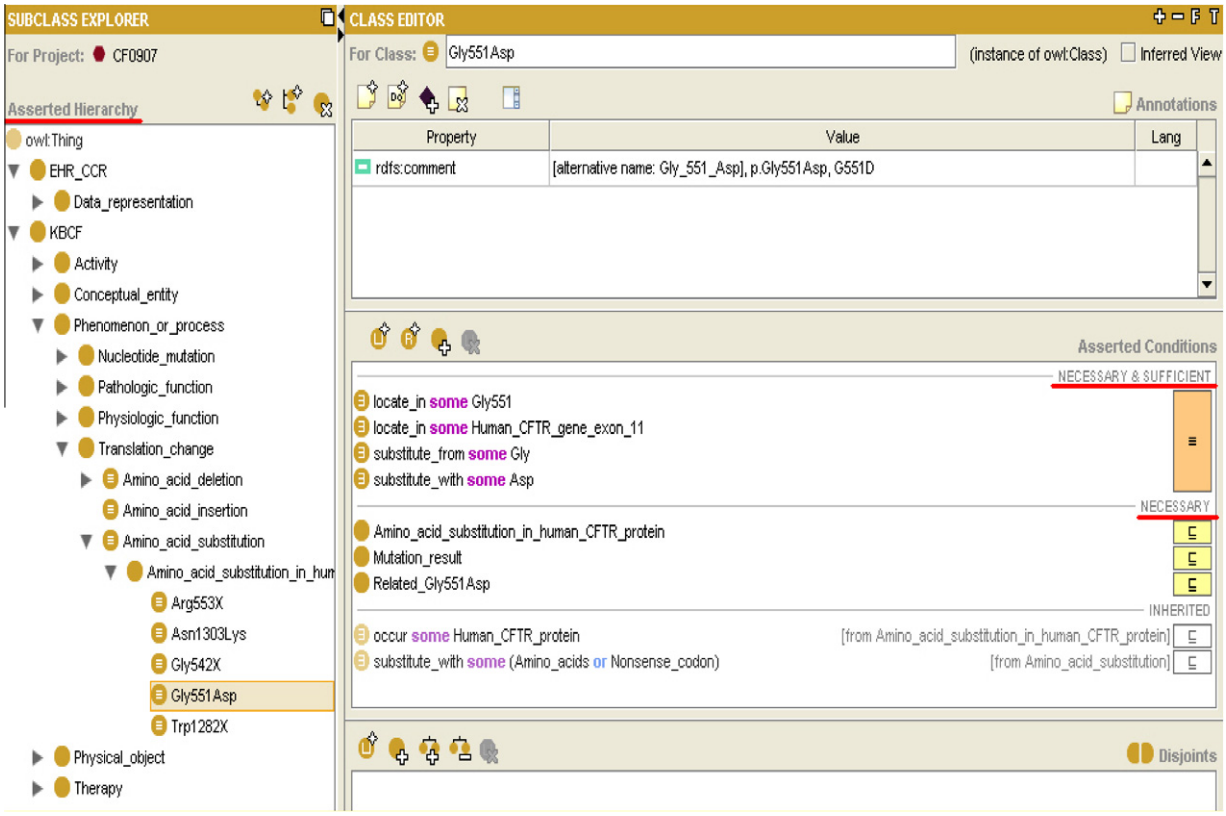

Fig. 1. Knowledge about the Gly551Asp mutation (also referred to as CFTR G551D) as represented in Protégé-OWL [12]. (In the hierarchy, Gly551Asp is a subclass of "amino_acid_substitution_in_human_CFTR_protein" and this means Gly551Asp is an amino acid change; under the Necessary and Sufficient conditions, "locate_in" and "substitute_with" describe the mutation position and result. For a detailed explanation of the knowledge representation, see Ref. [8].)

(3) Permanence: Most genetic data indicate permanent data about a patient, unlike most traditional laboratory test results that indicate transitory data.

(4) Reinterpretation: Although the results of genetic tests are permanent, the questions they answer vary over time as new research reveals linkages between genetic data and diseases.

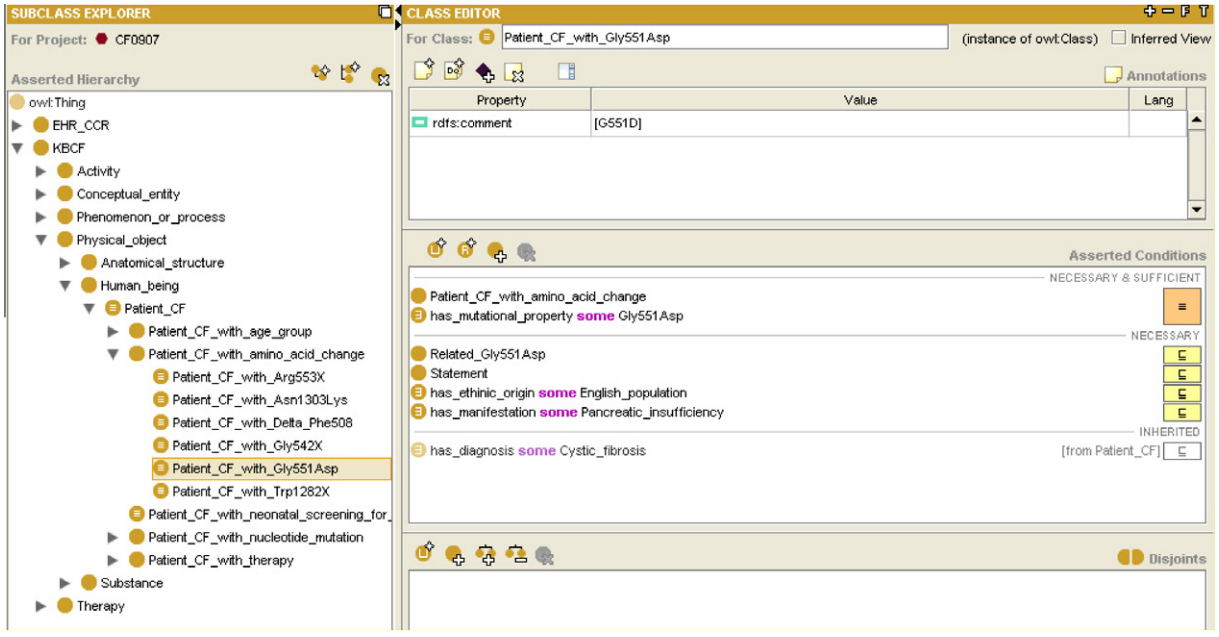

Fig. 2. Descriptions (health knowledge) of patients with Gly551Asp as a group. (Note that "Pancreatic Insufficiency" is represented as a Necessary Condition.)

The genetics knowledge in OntoKBCF is represented with the nomenclature for sequence variations prepared by den Dunnen [15]. We find that this widely accepted nomenclature maps well to the kinds of genetic data we are including in our prototype.

2.3. Continuity of Care Record

Semantic interoperability between systems, including clinical information systems and electronic health knowledge resources, has been identified as a significant challenge for clinical system development [16,17]. A standards-based health record architecture that can support semantic interoperability is a likely vehicle for integrating genetic data into future clinical applications. Our purpose in this research is to demonstrate our approach rather than to compare different available standards to proving superiority of one over others. We therefore selected the Continuity of Care Record (CCR) from the ASTM (formerly, the American Society for Testing and Materials) as an example to serve as the data structure underlying our prototype application.

The CCR is described as "a core data set of the most relevant administrative, demographic, and clinical information facts about a patient's healthcare, covering one or more healthcare encounters" [18]. We selected the CCR as the basis for the prototype because it is an established standard that might be used by others wishing to explore the integration of genetic data into EHRs. The CCR is based on a set of predefined headings and subheadings for structuring content, which simplified the development of our prototype. The headings and subheadings are text labels that indicate the meaning of subsequent text. CCR headings (labeled as "sections" in the specification) used in this research include *Problems*, *Results*, *Procedures*, *Patient* and *Date/Time* [18].

Data are included in a CCR document as a collection of "data objects" organized under the headings and subheadings and described with "attributes". For example, laboratory test results are included under the "Results" section, with each laboratory test result as a discrete data object, labeled with a "<Result>" tag. The various elements of these data objects include an identifier ("CCDataObjectID"), the "Type" of the test (such as "Hematology"), the name and code for the test (as a "Description"), the specimen (as a "Substance") and then one element for each of the individual results of the test (as one or more "Test" elements). The individual results each have their own "Description", "TestResult", and "NormalResult" elements which, in turn have elements of their own. For example, TestResult has the elements "Value" and "Units".

3. Methods

In order to demonstrate the feasibility of integrating genetic data into an EHR and then using the knowledge base to identify relevant genetics knowledge and health knowledge that can be used to assist with understanding and clinical decisions, we extended the CCR model so that it could support the representation of genetic data and then created a proof-of-concept EHR prototype, employing only those features necessary for delivering relevant information. We then created a data set of typical clinical cases and tested the ability of the prototype to filter, in a case-specific manner, relevant genetics knowledge and health knowledge from OntoKBCF.

3.1. Extension of the CCR

For the purposes of demonstrating our approach, we selected those parts of the CCR that would be relevant to the integration of *clinical data* and genetic data into a single user interface. For example, certain demographic data have particular importance since they are related to the genetics knowledge and can be used to select specific facts (such as possible conditions and recommended tests) in a context-sensitive manner.

We then considered the kinds of genetic data that might appear in the patient record, especially those relevant to the genetics knowledge. We examined the CCR to identify where genetic data would be included and the various subheadings and attributes that would need to be added to the CCR in order to represent these data appropriately. Part of our consideration was the inclusion of information to assist with interpretation of the genetic data [13].

3.2. Development of the EHR system prototype

The following factors were considered throughout the development of the system prototype: (1) this is proof of concept research so only a prototype was required; (2) the focus was to be on clinical and demographic data rather than administrative or financial data; (3) the prototype should harness the knowledge within OntoKBCF; (4) a specific clinical consultation scenario (limited consultation time and users who are not molecular genetics specialists) would be used to demonstrate feasibility; and (5) the basic interactions between an EHR user (e.g. a physician) and the EHR during a patient visit would be considered [19] (for example, an EHR user should be able to browse and edit patient's record, and get knowledge facts relevant to an individual patient from OntoKBCF via the EHR interface).

The two major parts of the prototype are the user interface and the underlying database. We designed the EHR portion of the database, which contain the clinical data and genetic data, to follow the organization of the CCR headings, subheadings and elements. We also included tables in the database to represent the health knowledge and genetics knowledge from OntoKBCF. The OntoKBCF facts are organized into tables based on the type of patient data used for filtering them. VB.NET, MS Access 2003 and Windows XP were used to construct the prototype.

3.3. Filtering knowledge for personalized information

As indicated in the Section 1, this research seeks to show that genetics knowledge and health knowledge from OntoKBCF can be personalized (i.e. customized), so that only the subset of knowledge facts relevant to a particular patient is displayed via the EHR system prototype. We developed a filter that used both patient's data and the knowledge in OntoKBCF to select a subset of knowledge facts for display in the EHR. The filter relies on patient-specific data to carry out its inferencing. For example, OntoKBCF includes facts that relate particular gene variants to a patient's ethnicity. The filter would consider displaying this fact if the patient in question is of that particular ethnicity.

When a patient's record is loaded, several tests are conducted: (1) if there is any knowledge fact in OntoKBCF that is related with this particular patient, i.e. filtering characteristics, then conduct the second check: (2) has this information already been included in the patient record? If not, then display the piece of information.

The filter itself is in the form of a collection of logic rules, similar to the rules typically found in clinical decision support systems. The rules use general, rather than fact-specific arguments. For example one rule expressed the logic: "if a patient's record has [gene] then search and display [gene related contents] when loading this patient's record". Both [gene] and [gene related contents] are groups of items, so this rule will be triggered whenever a single item in [gene] group is included in a patient's record and the record is loaded. The rules can therefore operate on new facts as they are added to the knowledge base. Of course, if new *types* of relationships are added to the knowledge base, new rules will be needed to make use of the facts that include the new relationships. For

the purposes of our prototype, we develop rules sufficient for the facts currently contained in OntKBCF.

3.4. Test data

In order to demonstrate the manner in which genetic data could be included in an EHR and be used to filter knowledge, a set of anonymous records were constructed. We endeavored to create a set that (1) used realistic patient parameters, (2) included a sufficient number of records to adequately test the filtering mechanism, (3) used several different mutation types, and (4) was relevant to the widest possible range of knowledge facts from OntoKBCF.

The patient records were derived from case reports found in PubMed. We selected those cases describing mutation types that corresponded to ones in OntoKBCF. Where several case reports included the same mutation type, we selected the one with the most detailed clinical description.

The selected cases provided data in narrative text form. In each case, we found it necessary to represent the content (age, sex, ethnicity, CFTR mutation types and other clinical aspects such as symptoms, diagnostic procedures and medications) according to the terminology used in OntoKBCF. Most of the case reports lacked complete details and it was therefore also necessary to add some hypothetical (but nevertheless realistic) data such as patient name and date of birth. Some of these hypothetical data (e.g. mutation results such as "G551D", and patient problems such as "pancreatic insufficiency") were added to each test record to ensure that the majority of relevant knowledge facts in OntoKBCF could be tested.

3.5. Evaluation

The evaluation of our prototype EHR consisted of loading each test record into the system and observing the facts that were selected for display from OntoKBCF. We hypothesized that the use of specific patient data would result in a significant reduction in the number of facts to be considered and that the subset of facts would vary from case to case.

4. Results

4.1. Extension of the CCR

Examination of the CCR identified several existing headings that would be needed for the clinical data in our cases, including "Actors" (which includes patient identifying information, such as name and date of birth, as well as demographic information, such as sex, ethnicity and age group), "Problems" (which includes patient symptoms and conditions), "Procedures" (which includes diagnostic and therapeutic procedures performed), and "Results" (which includes the results of procedures).

The heading "Results" is the appropriate location for including the genetic data for our prototype; therefore, we did not find a need to propose additional CCR headings. The CCR specifies "Types" for the test results under the "Results" heading for various types of laboratory testing, such as "Chemistry" and "Hematology". None of the existing types cover genetic testing, so we added "Molecular Genetics". Under this new type, we have specified "Gene Mutation Test" as a test class that can include terms for the tests that identify the variants of specific genes.

As we examined the types of results provided by gene mutation tests, we realized that the actual reported results (such as notation for a particular gene mutation) would be relatively meaningless to the clinician without sufficient molecular genetics training [13]. For example, understanding the meaning of "G551D" requires knowing that the notation refers to an amino acid sequence, as opposed to a nucleotide sequence, and that the "551" refers to a position in the sequence. We therefore further extended the CCR model to include two new attributes ("Molecular Structure" and "Molecular Position") to be used to characterize the results of gene mutation tests. Fig. 3 shows how the CCR has been extended to represent the result of a particular gene mutation test (in this case, CFTR Gene Mutation Test, a molecular genetic test that reports a single result).

4.2. Development of the EHR system prototype

The design of the EHR prototype is shown in Fig. 4. Clinical data and genetic data are stored in tables that follow the CCR data model (headings, data objects and attributes), while health knowledge and genetics knowledge are stored in tables that follow the OntoKBCF data model (facts represented as relationships between patients attributes and diseases attributes). Like any EHR user interface, the prototype user interface displays all of the patient-specific data (clinical data and genetic data) from the EHR tables. In addition, the interface displays a subset of OnfoKBCF knowledge (health knowledge and genetics knowledge). The subset uses patient-specific information to filter the knowledge: only the knowledge that is relevant to the particular patient will be displayed through the interface. Fig. 4 depicts the relationship between the patient data, knowledge, filtering mechanism and user interface.

The filtering mechanism makes use of the facts contained in the "OntoKBCF tables" (Fig. 4). These tables are primarily constructed automatically from an external knowledge base (OntokBCF). Thus, while the maintenance of the knowledge base may or may not be difficult, the knowledge in our prototype is easily updated, independent of external maintenance efforts.

4.3. Filters for personalized information

We developed a set of filter rules based on the patient's demographic data (including age, sex, and ethnicity) and the patient's genetic data and clinical data (including CFTR mutation type, and previous diagnostic and therapeutic procedures). Certain conditions are only relevant to a particular sex or age (e.g. symptoms of reproductive system for adolescent or adult patients), while certain mutation types may be relatively highly prevalent among certain ethnicities. The filter was designed to select for display those genetics knowledge and health knowledge facts that meet the following conditions: (1) relevant to a patient's specific age, sex and ethnic group, (2) relevant to specific diagnostic or therapeutic procedures in the patient's record, and (3) the knowledge facts that have not been already included in the existing patient's record (for example, if the patient has already developed pancreatitis, the fact that the patient's mutation predisposes to pancreatitis would not be included because "pancreatitis" is in his/her "diagnostic results" list.[2]

4.4. Test data

Our initial search for case reports returned 24 cases. After removing cases with duplicate mutation types, five cases remained, with CFTR gene mutations ΔF508, R117H, AA2183-G, Asn1303Lys, G1717-1A. Information was added to each case to

[2] Note that the purpose of the filter is to reduce information overload; suggesting that the patient with a particular gene is at risk for pancreatitis is of less importance in a patient who has already developed pancreatitis. The gene-condition fact could still be used by some other rule that might, for example, be used to suggest a genetic basis for an existing condition; we did not include this particular rule in our prototype.

```
<Results>
    <Result>
        <CCRDataObjectID>----</CCRDataObjectID>
        <Type><Text>Molecular Genetics</Text></Type>
        <Description>
            <Text>CFTR Gene Mutation</Text>
            <Code>
                <Value>C0805580</Value>
                <CodingSystem>UMLS</CodingSystem>
                <Version>2009AB</Version>
            </Code>
        </Description>
        <Substance>
            <Text>Venous Blood Substance</Text>
            <Code>
                <Value>C0229667</Value>
                <CodingSystem>UMLS</CodingSystem>
                <Version>2009AB</Version>
            </Code>
        </Substance>
        <Test>
            <Description>
                <Text>CFTR Gene</Text>
                <Code>
                    <Value>C1413365</Value>
                    <CodingSystem>UMLS</CodingSystem>
                    <Version>2009AB</Version>
                </Code>
            </Description>
            <TestResult>
                <Value>CFTR G551D</Value>
                <Molecular Structure>Amino Acid Sequence</Molecular Structure>
                <Molecular Position>CFTR Amino Acid Sequence 551</Molecular Position>
                <Units><Unit>Amino Acid</Unit></Units>
            </TestResult>
            <NormalResult>
                <Normal>
                    <Value>CFTR G551G</Value>
                    <Molecular Structure>Amino Acid Sequence</Molecular Structure>
                    <Molecular Position>CFTR Amino Acid Sequence 551</Molecular Position>
                    <Units><Unit>Amino Acid</Unit></Units>
                </Normal>
            </NormalResult>
        </Test>
    </Result>
</Results>
```

Fig. 3. A portion of an example CCR record, in XML, depicting one proposal for how genetic data can be included in the "Results" section of the record. [Elements in **bold-underline** font are the additions proposed in this research. As described in the text, we propose a new result type called "Molecular Genetics" which includes gene mutation tests, such as the CFTR Gene Mutation test. This test reports a single result – the CFTR gene – which is expressed with a gene annotation such as "CFTR G551D". Because of the difficulty that clinicians may have with understanding such annotations, we extend the CCR specification to include two additional attributes for the reported result: Molecular Structure (which in this case has the value "Amino Acid Sequence") and "Molecular Position" (which in this case has the value "CFTR Amino Acid Sequence 551"). Note that these two attributes also apply to the "NormalResult" which, in this case, has a newly specified value "CFTR G551G". We also add a new value ("Amino Acid") for the "Units" attribute (which is already included in the CCR standard). For clarity, some standard CCR subheadings, such as "DateTime" and "Source" have been omitted.]

give an explicit value (or make explicit the absence of a value) for 36 patient facts (demographics, symptoms, mutation types, screening tests, therapy, etc.) to which the filter would be sensitive. Additional patient information, not relevant to the filter, was excluded from the cases.

4.5. Evaluation

The test patients' records demonstrate that the filters are effective in reducing the number of facts (genetics knowledge and health knowledge) displayed and for tailoring those facts to suit patient parameters. Candidate facts are presented and can be selected and entered into the patient's record. The system is dynamic – retrieval and presentation are triggered automatically whenever the patient's record is loaded.

The following scenario is used to show the filtering function:

Jean, male, 12/11/1932, Italian; has cystic fibrosis.

In this scenario the *candidate* facts, i.e. health knowledge and genetics knowledge on the EHR interface from OntoKBCF include: 'bronchiectasis', 'deferred puberty', 'sterility'; related mutations include Asn1303Lys, AA2183-G, G1717-1A.

Table 1 presents two similar records, one for a male patient drawn from the case records and one for a hypothetical female version of the male patient. The table shows the OntoKBCF facts that have passed through the filters. The Age/Sex related health knowledge and genetics knowledge are different because of the different sex. This demonstrates the use of 'sex' (i.e. demographic data) as a filter. Genetic data can also be used as a filter: if a patient record has the mutation AA2183-G then AA2183-G will not appear in the candidate fact (for health knowledge and genetics knowledge) box.

Table 2 shows a comparison of number of knowledge fact counts before and after filtering for patient Jean and the other four patients. As expected, the number of knowledge facts is decreased after filtering.

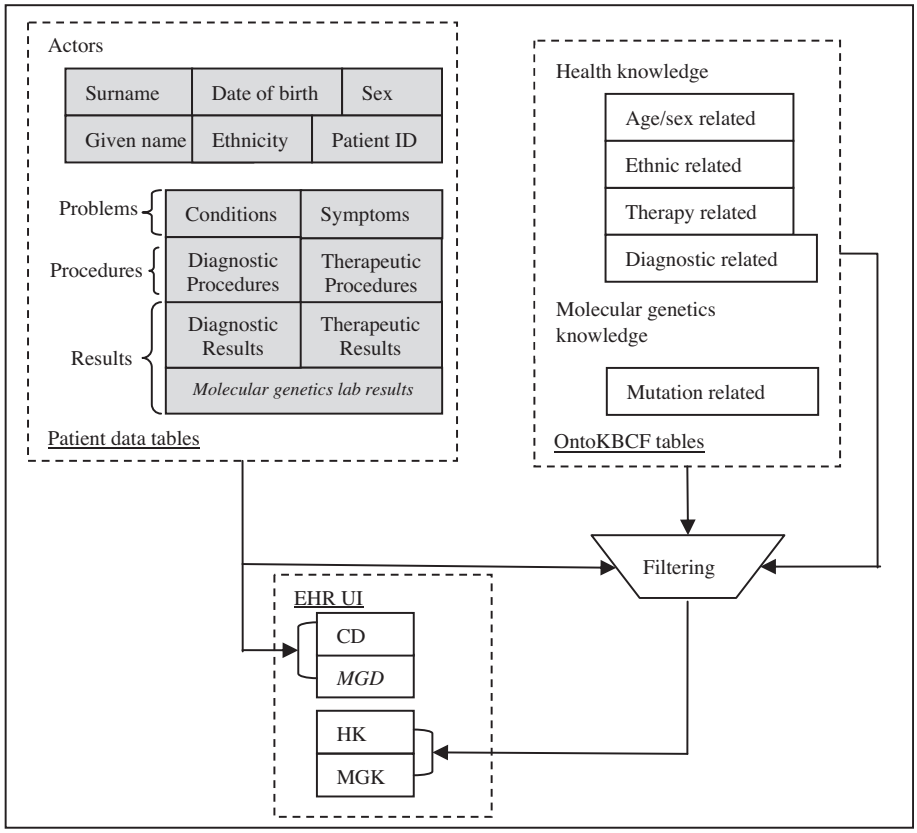

Fig. 4. Information components of the EHR system prototype, including patient data tables (which correspond to structures in the CCR), OntoKBCF tables, and the EHR user interface (UI). (Patient data from the EHR tables are displayed in the EHR UI, along with selected knowledge from the OntoKBCF tables. A filter applies knowledge from OntoKBCF itself to draw inferences, based on actual patient data, to select relevant knowledge facts for display in the EHR UI. CD, clinical data; MGD, molecular genetic data; HK, health knowledge; MGK, molecular genetics knowledge.)

Table 1
Different candidate knowledge facts (genetics knowledge and health knowledge) for opposite sex.

Patient names	Jean	Jean
Sex	Male	Female
Date of birth	12/11/1932	12/11/1932
Ethnicity	Italian	Italian
Age/sex related candidate facts	Bronchiectasis Deferred-puberty Sterility	Bronchiectasis Deferred-puberty Infertility and Scanty_cervix_mucus
Ethnicity related candidate facts	Asn1303Lys AA2183-G G1717–1A	Asn1303Lys AA2183-G G1717–1A
Diagnostic procedure related candidate facts	Nutrition_status_improved	Nutrition_status_improved

Note: the two records are identical except for sex – this shows the use of sex as a filter.

Fig. 5 is a screenshot of the EHR prototype showing a patient's data and relevant selected knowledge.

5. Discussion

This research has demonstrated, through the use of a set of test records, that a patient's genetic data, can be incorporated into patient records and that genetics knowledge and health knowledge (such as the content contained in OntoKBCF) can be tailored according to individual patient characteristics. This research sought to personalize genetics knowledge and health knowledge dynamically and automatically, in a simulated clinical system. The current CCR standard appears to require expansion in order to include genetic data; our methods can be an important reference for this type of expansion.

5.1. Knowledge representation

As previously described, [8] we considered a number of options for representing genetics knowledge in OntoKBCF. The GSVML (Genomic Sequence Variation Markup Language), for example, is a standard for communicating genomic information. The GSVML is a new emerging data exchange format that is useful in communications between health applications. However, in this research we focus on fitting genetic data into a standard EHR schema, as well as representing and customizing health knowledge and genetics knowledge; the establishment of a universal communication mechanism between the EHR and OntoKBCF is beyond the scope of this work. The current method for organizing health knowledge and genetics knowledge into an ontological-based knowledge base is satisfactory for our purpose of making the health knowledge and genetics knowledge accessible dynamically via an EHR. We chose a widely accepted nomenclature for sequence variations prepared by Dunnen [15] for representation of genetics knowledge in OntoKBCF. Conversion of the current genetics knowledge into another format (such as GSVML) could be readily accomplished by adding a separate layer on top of OntoKBCF.

5.2. Filters

Demographic data (age, sex and ethnicity), along with the patient's genetic data and clinical data, were used to filter (and customize) the full set of knowledge facts within OntoKBCF. Although OntoKBCF is not a comprehensive cystic fibrosis knowledge base yet, the role of the filter in tailoring genetics knowledge and health knowledge is still apparent. The data in Table 2 show that filter rules can significantly decrease the amount of knowledge facts presented to the user; in our cases, the reduction ranged from 90% to 94%. As the knowledge base becomes more comprehensive, many more knowledge facts will be available and filtering will become even more necessary.

Selection of filter rules is closely related to the available domain knowledge. One rule that was investigated but subsequently abandoned concerned lifestyle. Because of the relationship between cystic fibrosis and the respiratory system, there was an assumption that there would be some relationship between smoking and cystic fibrosis within bio-health sources [such as PubMed-Nucleotide, DNA Data Bank of Japan, The European Molecule Biology Laboratory (EMBL), On-line Mendelian Inheritance in Man (OMIM) and Cystic Fibrosis Mutation Database [20], Genetic Home Reference [21], Cystic Fibrosis Foundation [22], Mayo Clinic [23], two text books [24,25], and PubMed] and that the resulting information could be included in OntoKBCF, with lifestyle (specifically, smoking) used as a filter. However, there was insufficient information in the sources to make this possible, perhaps revealing a gap in biomedical research.

5.3. CCR headings

The CCR is a standard schema used for representing data in an EHR. However, there are no molecular genetics elements in the CCR and some headings for clinical data are not clear. We found it necessary to improvise in the interpretation and use of headings. As a type of laboratory result, the genetic data can share some attributes with traditional laboratory information (such as 'Substance' or 'Source'); however, accommodation of some of the specific characteristics of genetic data, such as amino acid position or nucleotide position, is difficult. We found that we needed to add more specific attribute tags to capture the corresponding data.

Although we added 'Mutation Results' to capture patients' sequence variation information as an example of results, it is too early to come to any general conclusion about the integration of genetic data into the CCR. Further investigations are required around integration into the CCR of a wider range of genetic data beyond the nucleotide and amino acid sequence variants used in this research.

5.4. Evaluation

The focus of the current research has been on the feasibility of bringing new types of patient-specific information and general knowledge together in an EHR in a new way. Our evaluation therefore focused on the ability of the prototype to bring patient-specific genetic data together with customized general genetic and health knowledge in an EHR environment. Additional evaluation is still needed in three areas: usability of the presentation approach, the accuracy of the facts presented based on specific patient data, and the scalability of the overall approach.

Given that the prototype is merely a vehicle to simulate the knowledge-delivery aspects of a real EHR, a usability study should focus on the cognitive aspects of the interaction [26], rather than such issues as navigation and layout. Specifically, the question of whether the context-specific delivery of knowledge enables a user to better understand the implications of genetic data to make more informed clinical decisions needs to be examined. Approaches such as those that have been used to study infobuttons (which deliver knowledge relevant to more traditional clinical data) [27] would be an appropriate reference for examining the usability of our prototype.

The accuracy of the facts presented by the prototype depends on the accuracy of two underlying contributors: the knowledge in OntoKBCF and the filtering rules. The former depends, in turn, on the accuracy of the resources (journal articles, textbooks and Cochrane reviews) used to create OntoKBCF, while the latter requires the filtering rules to perform correctly. Rather than attempting to evaluate these with two contributors separately, we believe the best approach will be to develop a gold standard for the knowledge facts (most likely using a panel of experts) that should be presented in each of some number of patient cases (whether real or simulated) and then measure the sensitivity and specificity of the prototype's selection of facts in each case. A root-cause analysis can be used to determine whether the fault for each false positive and false negative lies with OntoKBCF or the filtering rules.

An evaluation of the scalability of our approach will depend on (1) the availability of additional knowledge bases that cover genetic conditions other than cystic fibrosis and (2) a more comprehensive knowledge base of cystic fibrosis (OntoKBCF is not an exclusive cystic fibrosis knowledge base yet). Authoritative resources are available, such as the On-line Mendelian Inheritance in Man knowledge base [28]; their incorporation into an EHR such as ours will entail some increase in the number of filter rules, but we expect that many current ones will be reusable. We also expect that performance measures such as response time will not be adversely affected. However, the ability of our approach to filter large knowledge bases to produce sets of facts that will be manageable to the user is currently unknown, especially given the concerns about the rapid increase in genomic findings that are potentially relevant to individual patients but of unknown significance [29].

5.5. Comparison to previous work

To date, the exploration of merging a patient's genetic data into EHR, in any meaningful way, has been of a preliminary nature. The

Table 2
Comparison of facts counts before and after information filtering.

	Age/sex	Ethnicity	Mutation	Diagnostic procedure	Therapy	Total (%)
Before filtering	30	18	19	1	11	79 (100)
Subject: Jean	3	3	1	1	0	8 (10)
Subject: AB (Maria[a])	3	3	0	1	0	7 (9)
Subject: JB	3	2	0	0	0	5 (6)
Subject: AZ	3	2	0	0	0	5 (6)
Subject: Emily	2	1	0	0	2	5 (6)

Note: this table demonstrates the knowledge facts, including both genetics knowledge and health knowledge, have been filtered based on existing patient's data. The "fact counts" are the whole set of facts stored in database and each row with a fictitious patient's name is the facts counts that are displayed via the interface after filtering.

[a] AB is the original name from the case report and Maria is the name we used for test.

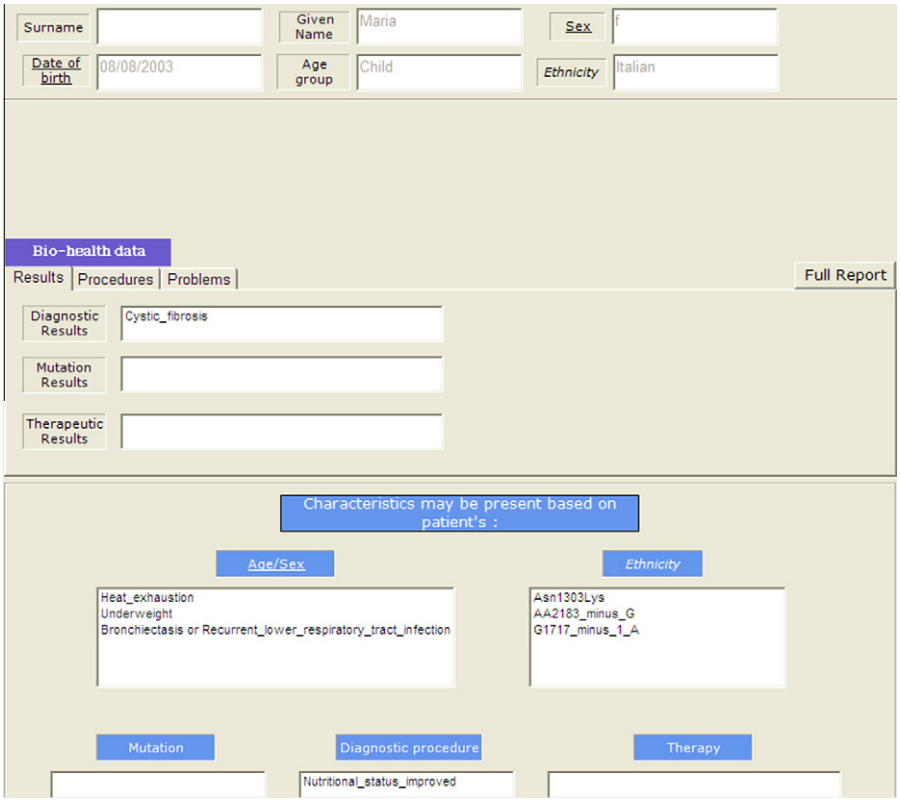

Fig. 5. A screenshot of the EHR prototype interface showing a patient's record and relevant knowledge (Demographic data are at the top of the screen, clinical data and genetic data are in the middle and titled as "Bio-health data", genetics knowledge and health knowledge are at the bottom of the screen and labeled as "Characteristics may be present based on patient's Age/Sex, Ethnicity, etc.").

GeneInsight Suite [30] is a platform at Partners Healthcare (Boston) that assists with the use of the results of molecular diagnostic testing in laboratories and in clinical environments. GeneInsight shares similar primary purposes with our prototype: to report DNA-based genetic testing results and to interpret the results. While GeneInsight has a much wider coverage, more comprehensive functions and has been implemented in multiple sites, our work breaks new ground in two ways: (1) we provide a method for exploiting external knowledge bases to drive the logic for providing data interpretation and (2) we integrate molecular genetic testing results into a simulated EHR environment.

Hoffman [7] recommended incorporating clinically significant genomic information into the Electronic Medical Record (EMR). However, his paper did not describe a method for accomplishing this. The description was more focused on challenges and available limited solutions. In contrast, our current research explores the organization of genetic data (mainly sequence variations) into a CCR-based EHR prototype, with emphasis on customization of genetics knowledge and health knowledge dynamically according to patient's data.

In a second study [31], Hoffman and colleagues provided clinical terminology that covered genetic concepts for use in a hospital information system (HIS). In our current research, the genetics knowledge and health knowledge were represented using the terminology of a knowledge base (OntoKBCF) which, in turn, was used for clinical decision support to actively and dynamically filter the knowledge facts displayed to the user. This filtering is triggered by default when a user interacts with the EHR. In Hoffman's research, terms were listed to describe findings by molecular diagnostic procedures and then imported into a commercial HIS system; with the terms serving as a dictionary for the domain of molecular diagnostic procedures. OntoKBCF, on the other hand is a solution in a totally different dimension, in that it connects the elementary concepts in molecular genetics and health fields, and it plays an active role, even after being embedded into an EHR system.

Murphy and colleagues [32] integrated clinical data and genetic data into a single architecture. However, the authors did not specify what genetic data had been included nor how the clinical data and genetic data had been integrated, making a comparison with our current work difficult.

The HL7 Clinical Genomics Working Group has established draft standards that can be used to send molecular genetic test results to EHRs [33]. The HL7 specification includes some aspects (such as reporting genetic variations) that are useful as reference information. However, the clinical genomics standards focus on explicit and correct transfer of information among systems or applications, and are not intended to provide all the attributes that would be needed for our purposes.

5.6. Limitations

The work presented here is limited in several ways. For the purpose of our study, we explored only one knowledge base, limited to a single disease. Cystic fibrosis is a well-studied genetic disease; as a result, many of the requirements for unambiguous names and clear relationships between genotypes and phenotypes have been addressed. Application of our approach to involve additional genetic diseases may not enjoy the same advantages. However, the types of information represented in OntoKBCF are not fundamentally different from information relevant to other diseases; our approach should therefore be applicable to other domains. OntoKBCF currently deals mostly with sequence variants, which represent only one type of molecular genetic data. Wider application of our

approach will require additional types of filter rules to address additional types of data, such as structural data or other types of variants, such as duplication, frame shift and complex changes.

Another limitation of our study relates to the minimal functionality of our prototype EHR. We addressed only a limited set of capabilities, and therefore any general conclusions should be drawn with caution. However, as a proof of concept, we believe the prototype is sufficient to demonstrate our ideas and that furthermore it provides a starting point and solid foundation for development and expansion of both the CCR standard and future EHR functionalities.

5.7. Significance

Our incorporation of genetic data into an EHR prototype is but a starting point on the path to true "personalized medicine". Studies such as ours will help both standards developers and systems designers on that path. The provision of health knowledge in the EHR context is not new, but the inclusion of molecular genetics knowledge breaks new ground for exploring ways to help clinicians cope with the large volume of strange new data that they will soon be encountering in patient care. Because the volume of molecular genetics knowledge is also large (and growing quickly), we also apply a novel method for filtering that knowledge, using a combination of the knowledge itself and patient-specific data to infer which facts will be useful to the clinician in a given context.

The research can be a starting point to build a comprehensive standards-based EHR system that offers personalized genetics knowledge and health knowledge. The standards-based prototype provides:

- A foundation for including molecular genetic data in personal health records.
- A data model that could be used as the basis for a range of future implementations.
- A potential platform for translational research to study the relationship between genetics knowledge and health knowledge.

The prototype, then, is a foundation for a personalized medicine framework and can also be used as a platform for translational research. For example, this prototype could be used to help study relationships between genotype and phenotype factors (including environmental factors) in individual clinical research subjects.

6. Conclusions

This study demonstrates one mechanism for incorporating patients' molecular genetic data into their electronic health records in a manner that facilitates personalized presentation of relevant genetic knowledge to support clinical decision making. Although a number of alternative representational schemes are possible, we show that at least one standard data model, the Continuity of Care Record, can be modified to support this task. Further research is needed to extend our approach to additional diseases and genetic data types in order to explore its generalizability to other extant knowledge bases and to define the necessary extensions of the CCR (or other suitable data models) to accommodate them.

Acknowledgments

This work had been supported by the Overseas Research Students Awards Scheme (UK), the University of Salford in the UK, and partly through intramural research funds from National Library of Medicine and the Clinical Center of the National Institutes of Health in the US. The authors thank Dr. Miao Sun for molecular genetics consulting and Dr. Yongsheng Gao for constructive discussions.

References

[1] US National Cancer Institute. Dictionary of cancer terms – personalized medicine; 2009. <http://www.cancer.gov/dictionary/?CdrID=561717> [cited 15.10.09].
[2] Ginsburg G, Willard H. Genomic and personalized medicine: foundations and applications. Transl Res 2009;154(6):277–87.
[3] Putre L. Personalized medicine. Getting genomic data into EMRs proves challenging. Hosp Health Netw 2009;83(7):20.
[4] Shortliffe EH, Perreault LE, Wiederhold G, Fagan LM. Medical informatics: computer applications in health care and biomedicine. 2nd ed. New York: Springer; 2001.
[5] Barnett GO, Barry MJ, Robb-Nicholson C, Morgan M. Overcoming information overload: an information system for the primary care physician. In: Fieschi M, Coiera E, Li Y-CJ, editors. MEDINFO2004. San Francisco (USA): IOS Press; 2004. p. 273–6.
[6] Sax U, Schmidt S. Integration of genomic data in electronic health records: opportunities and dilemmas. Methods Inf Med 2005;44:546–50.
[7] Hoffman MA. The genome-enabled electronic medical record. J Biomed Inform 2007;40:44–6.
[8] Jing X, Kay S, Hardiker N, Marley T. Ontology-based knowledge base model construction-OntoKBCF. In: MEDINFO 2007, Brisbane, Australia; 2007. p. 785–90.
[9] Jing X, Kay S, Hardiker NR. Designing a Bio-health information assistant within an EHR. In: Current perspectives in healthcare computing 2006. Healthcare computing 2006, Harrogate, England; 2006. p. 325–6.
[10] Bechhofer S, van Harmelen F, Hendler J, Horrocks I, McGuinness D, Patel-Schneider P, et al. OWL web ontology language reference-W3C recommendation, 10 February 2004. <http://www.w3.org/TR/owl-ref/> [cited 10.10.06].
[11] Zielenski J, Tsui LC. Cystic fibrosis: genotypic and phenotypic variations. Anu Rev Genetics 1995;29:777–807.
[12] Stanford Center for Biomedical Informatics Research. Protégé-the Ontology editor and knowledge acquisition system. <http://protege.stanford.edu/> [cited 2008 June 26th].
[13] Dequeker E, Stuhrmann M, Morris MA, Casals T, Castellani C, Claustres M, et al. Best practice guidelines for molecular genetic diagnosis of cystic fibrosis and CFTR-related disorders – updated European recommendations. Eur J Hum Genet 2009;17:51–65.
[14] Hirsch FR, Bunn PA. EGFR testing in lung cancer is ready for prime time. The Lancet 2009;10:432–3.
[15] den Dunnen J. Nomenclature for the description of sequence variations; 2007. <http://www.hgvs.org/mutnomen/> [cited 16.09.08].
[16] Semantic interoperability for better health and safer healthcare – research and deployment road map for Europe. European Communities; 2009. <http://www.eurorec.org/files/filesPublic/2009semantic-health-report.pdf> [cited 03.07.09].
[17] Kuhn KA, Wurst SHR, Bott OJ, Giuse DA. Expanding the scope of health information systems: challenges and developments. IMIA Yearbook Med Inform 2006;45(Suppl. 1):s43–52.
[18] ASTM. ASTM E2369-05. Standard specification for Continuity of Care Record (CCR); 2005.
[19] Van de Velde R, Degoulet P. In: Hannah KJ, Ball MJ, editors. Clinical information systems: a component-based approach. New York: Springer-Verlag; 2003.
[20] Hu Z, Mellor J, Wu J, DeLisi C. VisANT: an online visualization and analysis tool for biological interaction data. BMC Bioinform 2004;5(1):17.
[21] NLM. Genetics home reference – your guide to understanding genetic conditions. <http://ghr.nlm.nih.gov/> [cited 07.07.06].
[22] Cystic Fibrosis Foundation. What is cystic fibrosis? <http://www.cff.org/home/> [cited 06.07.06].
[23] Mayo Clinic Medical Services. Cystic fibrosis. <http://www.mayoclinic.com/health/cystic-fibrosis/DS00287/DSECTION=1> [cited 06.07.06].
[24] Harris A, Super M. Cystic fibrosis: the facts. 3rd ed. Oxford: Oxford University Press; 1995.
[25] Orenstein DM, Rosenstein BJ, Stern RC. Cystic fibrosis: medical care. Philadelphia: Lippincott Williams & Wilkins; 2000.
[26] Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004;37(1):56–76.
[27] Del Fiol G, Haug PJ, Cimino JJ, Narus SP, Norlin C, Mitchell JA. Effectiveness of topic-specific infobuttons: a randomized controlled trial. J Am Med Inform Assoc 2008;15(6):752–9.
[28] Schorderet DF. Using OMIM (On-line Mendelian Inheritance in Man) as an expert system in medical genetics. Am J Med Genet 1991;39(3):278–84.
[29] Kohane IS, Masys DR, Altman RB. The incidentalome: a threat to genomic medicine. JAMA 2006;296(2):212–5 [No abstract available. Erratum: JAMA 2006;296(12):1466].
[30] Aronson SJ, Clark EH, Babb LJ, Baxter S, Farwell LM, Funke BH, et al. The GeneInsight Suite: a platform to support laboratory and provider use of DNA-based genetic testing. Hum Mutat 2011;32:532–6.
[31] Hoffman M, Arnoldi C, Chuang I. The clinical bioinformatics ontology: a curated semantic network utilizing RefSeq information. Pac Symp Biocomput 2005;10:139–50.

[32] Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA annu symp proc 2006; 2006. p. 1040.
[33] HL7 Clinical Genomics Group. HL7 clinical genomics. <http://www.hl7.org/Special/committees/clingenomics/index.cfm> [cited 20.07.11].

Glossary

Clinical Data: patient-specific information traditionally found in a patient health record, including demographics, history, vital signs, symptoms, physical findings, and diagnostic test results

Continuity of Care Record (CCR): an ASTM standard for selecting and defining elements of a patient's health record

Cystic Fibrosis (CF): a genetic (autosomal recessive) disease in which the transportation of chloride ion across cell membranes is disrupted, leading to abnormal mucus secretions and subsequent complications in organs such as the lungs and pancreas

Cystic Fibrosis Transmembrane Conductance Regulator (CFTR): a protein (with an associated gene) that is responsible for transport of chloride ions across cell membranes; certain mutations in the gene responsible for coding this protein leads to cystic fibrosis

Electronic Health Record (EHR): a computer system that embodies patient health records in order to facilitate clinical data capture and display, as well as workflow processes such as clinical decision making

Health Knowledge: general (non-patient-specific) information about health and disease, including such aspects as etiology, diagnosis, prognosis and treatment

Genetic Data: refers to molecular genetic data in this paper; results of patient genetic tests that report various aspects of their genes and chromosomes, including locations and types of particular alterations in genetic sequence and named patterns of those variations (such as gene variants and mutations)

Genetics Knowledge: refers to molecular genetics knowledge in this paper; general (non-patient-specific) information about molecular genetics. In this paper the term particularly refers to molecular genetics knowledge represented in OntoKBCF

OntoKBCF: a particular knowledge base containing molecular genetics knowledge and clinical characteristics of cystic fibrosis, including basic concepts, their semantic relationships and descriptions of relevant knowledge facts

Personalized Information: customized data and knowledge that are selected based on a patient's clinical data, such as age, sex, and clinical characteristics, especially the results of molecular genetic testing