Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 40 (2007) 30–43
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information

Nicholas Sioutos [a], Sherri de Coronado [b,*], Margaret W. Haber [c], Frank W. Hartel [b], Wen-Ling Shaiu [d], Lawrence W. Wright [c]

[a] *Aspen Systems Corporation, USA*
[b] *National Cancer Institute Center for Bioinformatics, 6116 Executive Blvd., Suite 403, Bethesda, MD 20892-8335, USA*
[c] *National Cancer Institute Office of Communications, USA*
[d] *Management Systems Designers, Inc., USA*

Received 16 December 2005
Available online 15 March 2006

Abstract

Over the last 8 years, the National Cancer Institute (NCI) has launched a major effort to integrate molecular and clinical cancer-related information within a unified biomedical informatics framework, with controlled terminology as its foundational layer. The NCI Thesaurus is the reference terminology underpinning these efforts. It is designed to meet the growing need for accurate, comprehensive, and shared terminology, covering topics including: cancers, findings, drugs, therapies, anatomy, genes, pathways, cellular and subcellular processes, proteins, and experimental organisms. The NCI Thesaurus provides a partial model of how these things relate to each other, responding to actual user needs and implemented in a deductive logic framework that can help maintain the integrity and extend the informational power of what is provided. This paper presents the semantic model for cancer diseases and its uses in integrating clinical and molecular knowledge, more briefly examines the models and uses for drug, biochemical pathway, and mouse terminology, and discusses limits of the current approach and directions for future work.

Keywords: Biomedical vocabulary; Ontology development; Cancer research; Disease model; Cancer terminology

1. Introduction

Cancer is a genetic disease, in which a series of molecular events leads to runaway reproduction of cancer cells. Cancers are increasingly defined by their molecular origins and expressions, and the search for more effective treatments, screening, and prevention strategies now focuses on molecular mechanisms, expressions, and targets. Molecular information plays an ever more important role in a wide range of clinical activities: genetic tests to determine predisposition to cancer; serum protein assays, such as PSA, for cancer screening; a wide variety of molecular tests to help determine diagnosis and prognosis, and to measure response to treatment; identification of direct and indirect molecular targets for therapies; and identification of genetic and other patient characteristics leading to differences in drug response and side effects. The growing flood of new information—on tumors, patients, therapies, and techniques—is increasingly beyond what clinicians and researchers can handle with traditional approaches, in which individual clinical and research activities often proceeded in relative isolation, each with its own terminologies and information systems, and with fairly narrow—often paper-based—channels for communicating key information to others who might use it.

Over the last 8 years, the National Cancer Institute (NCI) has undertaken a major effort to integrate molecular and clinical cancer-related information within a unified

* Corresponding author. Present address: 113 Brookline St., Moraga, CA 94556, USA. Fax: +1 925 377 5960.
E-mail address: decorons@mail.nih.gov (S. de Coronado).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.02.013

biomedical informatics framework, with controlled terminology as its foundational layer. In 1997, the Enterprise Vocabulary Services (EVS) Project was launched, with the goal of integrating terminology used to code and retrieve information on all NCI-funded cancer research activities [1]. In 1999, the EVS was extended to include terminology needed for the NCI's Physician Data Query (PDQ®) databases and Web-based information services [2]. In 2000, the NCI Center for Bioinformatics (NCICB) started to create an NCI biomedical informatics infrastructure, and in 2003 the Cancer Biomedical Informatics Grid (caBIG) brought together NCI and some 40 Cancer Centers to develop a shared approach to coding, processing, and exchanging all types of cancer-related information [3]. By promoting the open and rapid exchange of information between a wide range of clinical and research systems, NCI hopes to greatly speed the development of new and more effective approaches to cancer, and at the same time tailor these approaches much more effectively to the individual characteristics of patients.

The EVS project created the NCI Thesaurus as a common reference terminology to underpin these efforts. Where earlier terminologies were designed for particular areas of activity—epidemiology, grant coding, clinical trial design, animal experiments, etc.—NCI Thesaurus is designed to represent and integrate information from these diverse areas, providing a structured and principled representation of key cancer-related concepts in areas such as cancers, findings, drugs, therapies, anatomy, genes, pathways, cellular and subcellular processes, proteins, and experimental organisms. Responding to user needs within a shared biomedical informatics environment, NCI Thesaurus has increasingly gone beyond the direct requirements of controlled terminology to create a model of how key concepts are defined and relate to each other, implemented in a deductive logic framework that can help maintain the integrity and extend the informational power of what is provided—shifting from being a controlled terminology toward what is today often called an ontology.

The need for unified controlled terminology has long figured as a grand challenge in biomedical informatics [4]. Intensive efforts have gone into collecting and cross-mapping terminologies already in use, most notably through the National Library of Medicine's Unified Medical Language System (UMLS) Metathesaurus [5], and attempting to move toward intercommunicating standards through such efforts as the US government Consolidated Health Informatics (CHI) and Federal Health Architecture initiatives [6]. Galen [7] and SNOMED-CT [8] are prominent examples of efforts to use deductive logic tools to provide more rigorously structured and richly informative biomedical terminology. At the same time, there has been a blossoming of biomedical ontologies attempting to more richly represent knowledge in particular fields in ways that often overlap with logic-based terminology efforts [9].

NCI Thesaurus attempts to combine features of both controlled terminology and ontology efforts: providing the logic-based reference terminology needed for reliable coding and cross-translation, and representing deeper knowledge in areas of particular importance to cancer research and clinical practice. The integration of molecular information in defining cancer concepts—in addition to older cellular, anatomic, morphologic, and clinical features—is the most advanced example of this shift towards ontological knowledge representation, and is explored in more detail below. Drug therapies, cancer prevention, and screening are also being linked to molecular targets and mechanisms, providing another important example of emerging conceptual models. Molecular concepts such as genes, proteins, pathways, oncogene deregulation, fusion protein expression, and chromosomal translocations, are important both as terminology for coding data and as conceptual links defining and relating other diagnostic and therapeutic concepts. Coverage of mouse models of cancer illustrates some of the unique challenges in creating a unified conceptual framework.

2. Materials and methods

The EVS Project is responsible for providing comprehensive resources that address NCI's terminology requirements. The NCI Thesaurus plays a central role in meeting these requirements, including providing terminology-based semantics for the NCI's caCORE biomedical informatics infrastructure [10]. These aspects of NCI Thesaurus' development and deployment are only briefly touched on here, as context for the semantic model. Its overall development, design, and purposes have been described elsewhere [11]. A detailed description of the description logic used to construct the NCI Thesaurus, the semantic relationships used in Thesaurus to relate concepts to each other, the formats in which NCI Thesaurus is distributed, and other technical details have also been published [12], as has the approach to managing change in the Thesaurus over time [13].

2.1. NCI Thesaurus in caCORE

The caCORE is an integrated suite of tools and resources supporting data management and application development, encompassing vocabulary, metadata, and biomedical data objects in a public domain technology stack [14]. NCI Thesaurus provides much of the semantics that underlie caCORE, and the caCORE provides real-time programming interfaces to access NCI Thesaurus and other EVS services. NCI employs caCORE widely, building numerous community portal sites, metadata contexts, and data repositories on it. Increasingly caCORE is being used to implement clinical and basic science resources at institutions external to NCI, such as the cancer center participants in the caBIG initiative [3].

In the current release of caCORE (3.0.1), EVS provides the "base semantics" for the metadata component and biomedical data objects. All metadata class and attribute

names correspond to concepts in NCI Thesaurus. Also, data objects have an attribute that points to the corresponding Thesaurus concept names and concept identifiers. Similarly, all the administered components from which metadata entities are constructed correspond to a concept name and concept identifier obtained from the EVS. This is what is meant by the term "base semantics": no matter where found, entities are named consistently; terms are bound unambiguously to concepts, and mean the same thing throughout caCORE.

This has an important implication for search of data repositories built on caCORE. The data accessible through an object that has an NCI Thesaurus concept code as an attribute can be considered as instances of the concept [12]. As such, the data have inherited the semantic associations between the concept of which they are an instance and the concepts that are related to it in the Thesaurus. By searching the data space for objects bearing the name or code of these related concepts, and then retrieving data through them, one can perform searches in the data space that are driven by information far richer than anything expressed in the object model itself. For example, at the time of this writing, the disease class in the caBIO UML model has two named relations to other classes; in contrast, in the Thesaurus, the named relationships between disease and other biomedical entities are over two dozen.

This integration also has important implications for NCI Thesaurus. As caCORE metadata and applications grow, NCI Thesaurus must grow with them, providing structured reference terminology for an ever widening array of efforts.

NCI Thesaurus can be accessed on the Web [15], through the published caCORE APIs [14], or by file download [16] in any of three formats: Ontylog XML, OWL Lite [17], and ASCII flat file. All NCI-developed caCORE components are distributed under open-source licenses that support unrestricted use by both non-profit and commercial entities.

2.2. Logic framework

The design philosophy and implementation of NCI Thesaurus has evolved to meet the needs of the NCI and its partners. It has also been shaped by issues of available technology and compatibility.

In its initial phases, NCI Thesaurus was a terminology system, and description logic was adopted to make creation and maintenance of the terminology easier, not to create a formal ontology. Description logic evaluation of assertions in the database, called *classification*, is used to detect logical inconsistencies in the terminology and to automatically organize concepts into related groups and hierarchies. This requires only the minimum set of semantic associations necessary to uniquely define concepts, and was economical and sufficient to meet NCI's operational requirements. Almost as a by-product, however, it resulted in a fairly expressive graph of semantic associations among basic science and clinical science concepts, and the terminology increasingly gained ontological properties.

The extensive modeling of disease concepts that is described in this paper represents the most radical departure from this previous design philosophy. This disease modeling follows a philosophy of including the maximum practical set of semantic relationships. Disease concepts are characterized by a rich set of relationships to molecular, genomic, proteomic, pharmaceutical, clinical, and biological concepts. These links are intended to support a variety of clinical and research uses, and the resulting model and descriptions come increasingly close to a full-blown ontology.

If the disease modeling is successful at providing support for clinical and research applications, then other areas covered by NCI Thesaurus may also be elaborated into ontological structures. Already, some initial steps have been taken in other areas, such as drugs, to meet growing demands in those areas.

NCI is committed to making the NCI Thesaurus an open terminology. No restrictions are placed on use of the Thesaurus, and its semantics and logical structure are non-proprietary. Others are encouraged to extend and reuse the terminology modeling as well as the terms. Toward that end, NCI Thesaurus is published in several formats, including two logic-based representations with somewhat different characteristics.

NCI Thesaurus is currently created in the Ontylog description logic (DL) system, the semantics of which are described in a recent paper [12]. Ontylog is not a highly expressive DL, but it includes some non-DL features and special support for terminology development. Although Ontylog DL is public domain, editors and other tools that implement the logic are largely or totally proprietary.

NCI relies on Web Ontology Language (OWL) format to make NCI Thesaurus modeling accessible to users of open, public domain terminology editors and related tools. To date, the semantic structures in NCI Thesaurus have mostly translated well from Ontylog into OWL [17]. The differences and issues, regarding such features as negation, complex role values, and non-description logic features, have to some extent been discussed elsewhere [12,17] and are raised in Section 4.

3. Results

NCI Thesaurus now includes some 43,000 biomedical concepts, separated into 20 logically distinct *kinds*. Kinds are similar to disjoint classes in description logics, but they also embody set and type qualities that are described more fully elsewhere [12]. These concepts are organized in multiple parent–child *is_a* hierarchies within each kind, as well as by over 100 distinct role relationships providing approximately 135,000 asserted and inherited logical links between pairs of concepts [18]. More than 10,000 of these concepts cover findings and disorders, and with most of the rest

being focused on molecular and biologic concepts or therapeutic and other agents. Molecular and clinical information are deeply interwoven in ways that can facilitate migration of "bench to bedside" insights in clinical care, hypothesis generation, molecular target identification, or other activities. The domain areas and sample use cases below make clear the relevance of NCI Thesaurus to these issues.

3.1. Disease model

The disease model is structured to support rigorous definition of cancers and other diseases, specifying distinguishing features and enforcing logical organization in their classification. These molecular, cellular, anatomic, morphologic, and clinical features provide an effective way to link molecular findings to cancers, identify diverse disease entities that share common molecular signatures as potential therapeutic targets, indicate the particular biologic events that characterize and often determine the outcome of a disease, and provide important information for researchers, health professionals, and the public. Where a disease is characterized by the absence of certain molecular, pathologic, laboratory, and clinical findings, this is expressed by use of exclusion roles in Ontylog description logic (translated into negations in OWL). Characteristics that are interdependent are linked together under specific role groups, signifying that they are being asserted as a unit; role groups have been particularly important in flagging certain aspects of newly emerging molecular features, and some issues with how they are used in this context will be discussed later.

3.1.1. Disease model relevance to diagnosis, prognosis, and treatment

Precise pathologic diagnosis, with correlations to prognostic parameters, is of paramount importance in the work-up of cancers. Clinical oncologists and patients need this information to select appropriate treatment options. Knowledge of associated molecular abnormalities can direct pathologists and clinical oncologists to submit tissue samples for molecular diagnostic work-up when a definitive diagnosis cannot be made, to confirm a diagnosis, or to help determine prognosis. Basic scientists can use the molecular abnormality-disease associations too, enabling them to link their research findings to related disease and molecular patterns.

The following examples of three lymphoid cancers illustrate how the role-based characterization of molecular identity, cellular origin, pathology, anatomy, and clinical course of these malignancies can support such clinical activities.

3.1.1.1. Gastric mucosa-associated lymphoid tissue (GMALT) lymphoma. This common, indolent extranodal marginal zone B-cell lymphoma is often associated with Helicobacter pylori infection, and is defined by the presence of centrocyte-like cells and lymphoepithelial lesions in the gastric mucosa [19,20]. It is characterized by the t(11;18)(q21;q21) translocation that involves the apoptosis inhibitor gene AP12 on chromosome 11 and the MLT (also known as MALT1) gene on chromosome 18. The t(11;18)(q21;q21) translocation results in the expression of the chimeric AP12-MLT fusion transcript [21]. The t(11;18)(q21;q21) translocation occurs in approximately 40% of gastric mucosa-associated lymphoid tissue lymphomas, and, when it is present, the lymphoma is resistant to Helicobacter pylori treatment [22].

This clinically relevant information is largely captured in the role-based associations shown below:

Molecular abnormalities:
Disease_May_Have_Cytogenetic_Abnormality: Trisomy 3
Disease_May_Have_Cytogenetic_Abnormality: Trisomy 18

Role group 1:
Disease_May_Have_Cytogenetic_Abnormality: t(11;18)(q21;q21)
Disease_May_Have_Molecular_Abnormality: AP12-MLT fusion protein expression

Histogenesis:
Disease_Has_Normal_Cell_Origin: Post-germinal center marginal zone B-lymphocyte

Pathology:
Disease_Has_Abnormal_Cell: Centrocyte-like cell
Disease_May_Have_Abnormal_Cell: Neoplastic monocytoid B-lymphocyte
Disease_May_Have_Abnormal_Cell: Neoplastic plasma cell
Disease_May_Have_Finding: Lymphoepithelial lesion

Anatomy:
Disease_Has_Primary_Anatomic_Site: Stomach
Disease_Has_Normal_Tissue_Origin: Gut associated lymphoid tissue

Clinical information:
Disease_May_Have_Finding: Indolent clinical course
Disease_May_Have_Associated_Disease: Hepatitis C

A role group indicates that the abnormal transcript is the result of the specific translocation. The clinical oncologist, alerted to this variable feature, could request a molecular work-up in the endoscopic biopsy or surgical specimen in question to determine the appropriate therapeutic approach. The t(11;18)(q21;q21) translocation is also linked to extranodal marginal zone B-cell lymphomas involving several other anatomic sites (e.g., bronchial mucosa-associated lymphoid tissue lymphomas), but not in nodal or splenic marginal zone B-cell lymphomas. Thus, the molecular characterization of

these diseases can be used effectively in diagnostic practice to differentiate between small, atypical lymphoproliferations that raise the question of marginal zone B-cell lymphoma.

3.1.1.2. Diffuse large B-cell lymphoma with an activated B-cell expression profile. This recently defined biologic subtype of diffuse large B-cell lymphomas (DLBCLs) has a unique molecular signature, and is characterized by the overexpression of CD44, BCL-2, Cyclin D2, IRF4/MUM1, and PKCbeta1 genes. This type of DLBCL originates from activated B-lymphocytes, is more often centroblastic than immunoblastic, and has an unfavorable prognosis. The following characteristics have been used to define it:

Molecular abnormalities:
Disease_Has_Molecular_Abnormality: BCL-2 messenger RNA overexpression
Disease_Has_Molecular_Abnormality: CD44 messenger RNA overexpression
Disease_Has_Molecular_Abnormality: Cyclin D2 messenger RNA overexpression
Disease_Has_Molecular_Abnormality: IRF4/MUM1 messenger RNA overexpression
Disease_Has_Molecular_Abnormality: PKCbeta1 messenger RNA overexpression
Disease_Has_Molecular_Abnormality: Increased NFkappaB pathway activation

Histogenesis:
Disease_Has_Normal_Cell_Origin: Activated B-lymphocyte

Pathology:
Disease_Has_Abnormal_Cell: Neoplastic large B-lymphocyte
Disease_May_Have_Abnormal_Cell: Neoplastic centroblast
Disease_May_Have_Abnormal_Cell: Neoplastic immunoblast
Disease_Has_Finding: Diffuse pattern
Disease_Excludes_Finding: Nodular pattern

Anatomy:
Disease_Has_Primary_Anatomic_Site: Lymphatic system
Disease_Has_Normal_Tissue_Origin: Lymphoid tissue

Clinical information:
Disease_May_Have_Finding: Unfavorable clinical outcome
Disease_May_Have_Finding: Lymphadenopathy

A second, distinct biologic DLBCL subtype has also been described based on a different gene expression profile (DLBCL with a germinal center B-cell expression profile). It is characterized by the overexpression of A-myb, BCL-6, CD10, and LMO2 genes. This subtype originates from germinal center B-lymphocytes, usually has centroblastic morphology and carries a favorable prognosis [23–25]. These two new biologic DLBCL entities represent an excellent example of the contribution of basic science to clinical medicine, by shedding new light on the diverse clinical outcomes of DLBCLs with apparently identical morphologies. A comparative review of molecular, cellular, and clinical models of DLBCL with an activated B-cell expression profile versus DLBCL with a germinal center B-cell expression profile could guide the testing of tissue samples from such cases for the appropriate biomarkers (Table 1). Precise pathologic diagnosis, based on molecular signatures, could suggest alternative therapeutic approaches for cases that express biomarkers linked to an unfavorable prognosis.

3.1.1.3. Childhood precursor B-lymphoblastic leukemia. This common pediatric leukemia is associated with specific chromosomal translocations. Each translocation, when present, gives rise to a specific fusion protein expression. The presence or absence of a particular translocation may define prognosis.

Table 1
DLBCL with an activated B-cell expression profile vs. DLBCL with a germinal center B-cell expression profile: a comparative view of the NCI Thesaurus molecular, cellular origin, and clinical models

Role	Diffuse large B-cell lymphoma with an activated B-cell expression profile	Diffuse large B-cell lymphoma with a germinal center B-cell expression profile
Disease_Has_Molecular_Abnormality	BCL-2 messenger RNA overexpression CD44 messenger RNA overexpression Cyclin D2 messenger RNA overexpression IRF4/MUM1 messenger RNA overexpression PKCβl messenger RNA overexpression Increased NFκB pathway activation	A-myb messenger RNA overexpression BCL-6 messenger RNA overexpression CD 10 messenger RNA overexpression LMO2 messenger RNA overexpression BCL-2 translocation REL gene amplification
Disease_Has_Normal_Cell_Origin	Activated B-lymphocyte	Germinal center B-lymphocyte
Disease_May_Have_Finding	Unfavorable clinical outcome	Favorable clinical outcome

Molecular abnormalities:
Role group 1:
Disease_May_Have_Cytogenetic_Abnormality: t(12;21)(p13;q22)
Disease_May_Have_Molecular_Abnormality: TEL-AML1 fusion protein expression
Disease_May_Have_Finding: Favorable clinical outcome

Role group 2:
Disease_May_Have_Cytogenetic_Abnormality: t(1;19)(q23;p13)
Disease_May_Have_Molecular_Abnormality: E2A-PBX1 fusion protein expression
Disease_May_Have_Finding: Unfavorable clinical outcome

Role group 3:
Disease_May_Have_Cytogenetic_Abnormality: t(4;11)(q21;q23)
Disease_May_Have_Molecular_Abnormality: MLL-AF4 fusion protein expression
Disease_May_Have_Finding: Unfavorable clinical outcome

Role group 4:
Disease_May_Have_Cytogenetic_Abnormality: t(9;22)(q34;q11)
Disease_May_Have_Molecular_Abnormality: p190 fusion protein expression
Disease_May_Have_Molecular_Abnormality: p210 fusion protein expression
Disease_May_Have_Finding: Unfavorable clinical outcome

Histogenesis:
Disease_Has_Normal_Cell_Origin: Precursor B-lymphoblast

Pathology:
Disease_Has_Abnormal_Cell: Neoplastic B-lymphoblast
Disease_May_Have_Finding: Extensive bone marrow and blood involvement

Anatomy:
Disease_Has_Associated_Anatomic_Site: Hematopoietic system
Disease_Has_Primary_Anatomic_Site: Lymphatic system
Disease_Has_Normal_Tissue_Origin: Lymphoid tissue

Clinical information:
Disease_May_Have_Finding: Bone pain
Disease_May_Have_Finding: Hepatomegaly
Disease_May_Have_Finding: Lymphadenopathy
Disease_May_Have_Finding: Splenomegaly

The specific translocations, fusion proteins, and clinical outcomes are grouped together in role groups. This detailed molecular information can be used by pathologists and hematologists to establish the molecular identity of a new patient, to monitor the response to treatment, and to detect minimal residual disease [26–38]. Other acute leukemias are also associated with the above translocations; for example, t(4;11)(q21;q23) is also linked to acute myeloid leukemia with 11q23 (MLL) abnormalities, and acute leukemia of ambiguous lineage.

3.1.2. Disease modeling for basic science

The impact of molecular biology in defining unique disease entities with distinct molecular signatures and diverse clinical courses has created the need for an expedited transfer of the molecular advances in cancer research to clinical medicine. Furthermore, the shared molecular pathways involved in the pathogenesis of various cancers raise the possibility of potential therapeutic intervention by single agents that may interfere with the specific molecular events. Systematic elucidation of these features of human cancers can greatly facilitate translational research.

Basic researchers working on cell lines or tissues from animal models or humans can link their genetic data to specific molecular events that in turn are associated with specific cancers. For example, the EWS gene (22q12.2) is linked through role relationships with chromosomal translocations and abnormal protein expressions that characterize a family of cancers grouped together under the term Ewing's sarcoma/peripheral primitive neuroectodermal tumors (Table 2) [39]. These molecular concepts are also associated with other cancers—morphologically unrelated to Ewing's sarcoma/peripheral primitive neuroectodermal tumors—such as myxoid extraskeletal chondrosarcoma and desmoplastic small round cell tumor (Table 2).

Detailed associations are defined between oncogene, tumor-suppressor gene, and mismatch repair gene abnormalities and cancers, including hereditary (familial) cancers and cancer-associated syndromes. In addition, precancerous conditions (e.g., endometrial intraepithelial neoplasia, pancreatic intraepithelial neoplasia) are linked to specific genetic abnormalities, underscoring the impact of certain molecular events in the transformation of dysplastic and in situ lesions to invasive cancers. Researchers studying a particular gene abnormality can immediately obtain information on a diverse list of common and infrequent cancers and syndromes that are associated with the particular gene of interest.

For example, a researcher who has detected amplification of KRAS oncogene in a cell line, animal model, or human tissue can directly retrieve the following cancers as associated with KRAS abnormalities: breast carcinoma, cholangiocarcinoma, colorectal carcinoma, head and neck carcinoma, lung carcinoma, and endometrial endometrioid carcinoma. Similarly, a researcher who has identified inactivation of BRCA2 tumor-suppressor gene can retrieve information on hereditary human cancers (hereditary female and male breast carcinoma, ovarian and fallopian tube carcinoma), and non-hereditary human cancers

Table 2
Cancers, chromosomal translocations, and fusion proteins associated with Ewing's sarcoma gene in the NCI Thesaurus disease models

Role group	Ewing's sarcoma/peripheral primitive neuroectodermal tumor	Myxoid extraskeletal chondrosarcoma	Desmoplastic small round cell tumor
Role group 1	Disease_May_Have Cytogenetic Abnormality t(11;22)(q24;q12) Disease_May_Have_Molecular_Abnormality EWS-FL1 fusion protein expression	t(9;22)(q22-31;q12) EWS-NR4A3 fusion protein expression	t(11;22)(p13;q12) EWS-WT1 fusion protein expression
Role group 2	Disease_May_Have Cytogenetic Abnormality t(21;22)(q22;q12) Disease_May_Have_Molecular_Abnormality EWS-ERG fusion protein expression		
Role group 3	Disease_May_Have Cytogenetic Abnormality t(17;22)(q21;q12) Disease_May_Have_Molecular_Abnormality EWS-E1AF fusion protein expression		
Role group 4	Disease_May_Have Cytogenetic Abnormality t(7;22)(p22;q12) Disease_May_Have_Molecular_Abnormality EWS-ETV1 fusion protein expression		
Role group 5	Disease_May_Have Cytogenetic Abnormality t(2;22)(q13;q22),t(3;18)(p21;q23) Disease_May_Have_Molecular_Abnormality EWS-FEV fusion protein expression		
Role group 6	Disease_May_Have Cytogenetic Abnormality t(1;22)(p36.1;q12) Disease_May_Have_Molecular_Abnormality EWS-ZSG fusion protein expression		

(ovarian carcinosarcoma, pancreatic carcinoma), and precancerous conditions (pancreatic intraepithelial neoplasia) that are associated with BRCA2 mutations (Table 3).

These molecular features can both define a specific subtype of a particular cancer or syndrome, and also point to other, associated diseases. For example, Turcot syndrome type 1, in addition to the MLH1 mismatch repair gene abnormality, may also carry abnormalities in the mismatch repair genes MSH2 and PMS2. Patients with type 1 variant of Turcot syndrome develop glioblastomas, and are at risk for hereditary non-polyposis colorectal cancer. Turcot syndrome type 2, however, is always characterized by abnormalities in the APC tumor-suppressor gene only, and patients with this variant develop medulloblastomas and adenomatous polyposis coli (Table 4) [40].

Distinguishing details about shared molecular events that characterize various cancers also constitute important information for researchers. For example, t(X;17)(p11;q25) and the associated ASPL-TFE3 fusion protein expression characterize a subgroup of renal cell carcinomas associated with Xp11.2 translocations/TFE3 gene fusions, as well as the morphologically and clinically unrelated alveolar soft part sarcomas. [39] However, the former cancer is associated with a balanced chromosomal translocation, whereas the latter is associated with an unbalanced translocation (Table 3). This latter information is linked to the chromosomal translocations through use of role groups.

3.1.3. Linking the disease and drug models

The NCI Thesaurus disease model represents a prototype that can serve as the basic framework to link diverse cancers that share a unique molecular abnormality to specific drugs known to interfere with the molecular pathways involved in the pathogenesis of such cancers (targeted therapy). Roles are being created in the NCI Thesaurus to link drug mechanisms of action to molecular pathways, which

Table 3
Molecular abnormalities linked to representative precancerous conditions, cancers, and cancer-related syndromes in the NCI Thesaurus disease models

Molecular abnormality	Linked condition
KRAS gene amplification	Breast carcinoma Cholangiocarcinoma Colorectal carcinoma Head and neck carcinoma Lung carcinoma Endometrial endometrioid carcinoma
PTEN gene inactivation	Hepatocellular carcinoma Invasive breast carcinoma Acinar prostate adenocarcinoma Follicular thyroid carcinoma Glioblastoma Pancreatic adenocarcinoma Type 1 endometrial adenocarcinoma Endometrial endometrioid adenocarcinoma Endometrial intraepithelial neoplasia Ovarian endometrioid adenocarcinoma Ovarian mixed epithelial neoplasm Mycosis fungoides Cowden syndrome
BRCA2 gene inactivation	Hereditary female breast carcinoma Hereditary male breast carcinoma Hereditary ovarian carcinoma Hereditary fallopian tube carcinoma Ovarian carcinosarcoma Pancreatic carcinoma Pancreatic intraepithelial neoplasia
MLH1 gene inactivation	Endometrial carcinoma Gastric carcinoma Hereditary ovarian carcinoma Turcot syndrome type 1
t(X;17)(p11;q25) ASPL-TFE3 fusion protein expression	Renal cell carcinoma with t(X;17)(p11;q25) [balanced chromosomal translocation] Alveolar soft part sarcoma (unbalanced chromosomal translocation)

Table 4
Turcot syndrome and subtypes: molecular characteristics and associated diseases in the NCI Thesaurus disease model

Role	Turcot syndrome type 1	Turcot syndrome type 2
Disease_May_Have_Molecular_Abnormality	MLH1 gene inactivation MSH2 gene inactivation PMS2 gene inactivation	
Disease_Has_Molecular_Abnormality		APC gene inactivation
Disease_Has_Associated_Disease	Glioblastoma	Medulloblastoma Familiar adenomatous polyposis
Disease_May_Have_Associated_Disease	Hereditary non-polyposis colon cancer	

may define or be associated with cancers. The integration of disease and drug models in the NCI Thesaurus will serve as a reference for correlating molecular abnormalities to potential new therapies.

For example, C-KIT (CD117) tyrosine kinase protein, a member of the type III receptor tyrosine kinase family, is involved in the pathogenesis of several cancers [41]. When C-KIT tyrosine kinase protein is overexpressed, it is capable of protecting tumor cells against apoptosis, thus promoting tumor growth [42]. Tumors known to have associated C-KIT tyrosine kinase protein overexpression are molecularly modeled in the NCI Thesaurus, and some of them are characterized by C-KIT gene mutations as well (Table 5). C-KIT molecular abnormalities can potentially serve as effective targets for small molecule inhibitors, such as imatinib mesylate (STI-571, Gleevec) which induces apoptosis and inhibits tumor cell proliferation [41]. The therapeutic potential of imatinib mesylate has been demonstrated in specific tumors characterized by C-KIT abnormalities (gastrointestinal stromal tumors, small cell carcinoma cell lines, and neuroblastoma cell lines). NCI Thesaurus is being extended to provide researchers and clinical oncologists with authoritative information on drugs such as imatinib mesylate and all cancers characterized by C-KIT and other pathway abnormalities.

3.2. NCI Thesaurus drug model

The NCI Thesaurus drug terminology currently contains approximately 4000 single agents and over 3000 combination therapies for cancer drugs in clinical treatment and prevention trials, or used for supportive care. Agents are incorporated into Thesaurus from multiple primary research and clinical treatment related data sources, and are used by a broad array of applications. Drugs are classified on the basis of functional, structural, and therapeutic intent hierarchies, if possible, with text definitions and computable role relationships for mechanism of action, physiologic effects, known effects on gene products as molecular targets, and affected anatomic structures, including subcellular targets, if applicable. Therapeutic intent, including food and drug administration (FDA) or standard usage, as well as clinical trial drug uses, are represented as properties(facts asserted directly for a single concept) rather than as roles, to avoid possibly inappropriate inheritance of these assertions at lower level nodes of the drug hierarchies.

Table 5
NCI Thesaurus cancer models characterized by C-KIT abnormalities

Molecular abnormality	Associated condition
C-KIT gene mutation and C-KIT tyrosine kinase protein overexpression	Gastrointestinal stromal tumor Dysgerminoma Seminoma Mastocytosis
C-KIT tyrosine kinase protein overexpression	Neuroblastoma Small cell lung carcinoma Large cell lung neuroendocrine carcinoma

Drug concepts include extensive synonymy, designed to maximize retrieval of the correct drug concept using a wide variety of naming conventions; these include generic, chemical, and brand names, abbreviations, acronyms, research codes, and other identifiers such as CAS, NSC, IND, and FDA unique ingredient identifier (UNII) codes. Combination therapy concepts link back, through role relationships, to each component single agent concept. Elements of the drug model are designed to be compatible and shareable with other public, freely available drug data information resources made available as standards by the FDA, the Department of Veteran's Affairs (VA), and the National Library of Medicine (NLM), including VA's National Drug File Reference Terminology (NDF-RT) and NLM's RxNorm [43].

The drug model can be used to support targeted drug research, as mentioned above, but also to support pharmacokinetic and pharmacogenomic research. For example, the heterogeneity in chemotherapy responses observed across patient populations can be the result of genetic polymorphisms in genes that govern drug metabolism and disposition. These polymorphisms may alter the efficacy of the chemotherapy agents and the likelihood of an adverse reaction. Semantics to describe the inter-relationships between normal genes, alleles, gene products, drug metabolism, and drug agents are therefore needed to facilitate both research and clinical care.

For example, a patient with non-small cell lung cancer fails to respond to other types of chemotherapy. The patient's physician wants to know whether Iressa, the

current approved drug for this condition, has any variation in metabolism or response rates based on genetic heterogeneity. Fig. 1 shows the drug information model in NCI Thesaurus that supports retrieval of this information. Iressa is linked to the concept "enzyme interaction" in the mechanism of action hierarchy. It is seen to have a physiologic effect of "tyrosine kinase inhibition" and a molecular target of "epidermal growth factor receptor." In turn, "epidermal growth factor receptor" is encoded by "EGFR gene," and "EGFR gene variant" is a variant of "EGFR gene." One of these EGFR gene variants affects the efficacy of Iressa [44,45]. Therefore, the physician might choose to do a genetic test to identify whether the patient has this variant.

Most drug-metabolizing enzymes exhibit clinically relevant genetic polymorphisms. For example, the cytochrome P450 family of enzymes exhibit common allelic variations in the population that affect drug metabolism in the patient (see Table 6). One member of the family, the CYP2D6 enzyme, metabolizes numerous clinical drugs, including beta blockers, antidepressants, and antipsychotics. So far, about 80 CYP2D6 variants have been identified. CYP2D6*2 has enzymatic activity that is comparable with the wild-type isoform, but this allele is associated with allelic duplication or amplification in high frequency. Many people who are ultra-rapid metabolizers have been found to have multiple copies of functional CYP2D6*2 variants. Therefore, individuals carrying CYP2D6*2 may have very high enzymatic activity and require unusually high doses of drugs metabolized by this enzyme to maintain therapeutic concentrations. Conversely, the CYP2D6*10 variant, found with high frequency in Asian populations, results

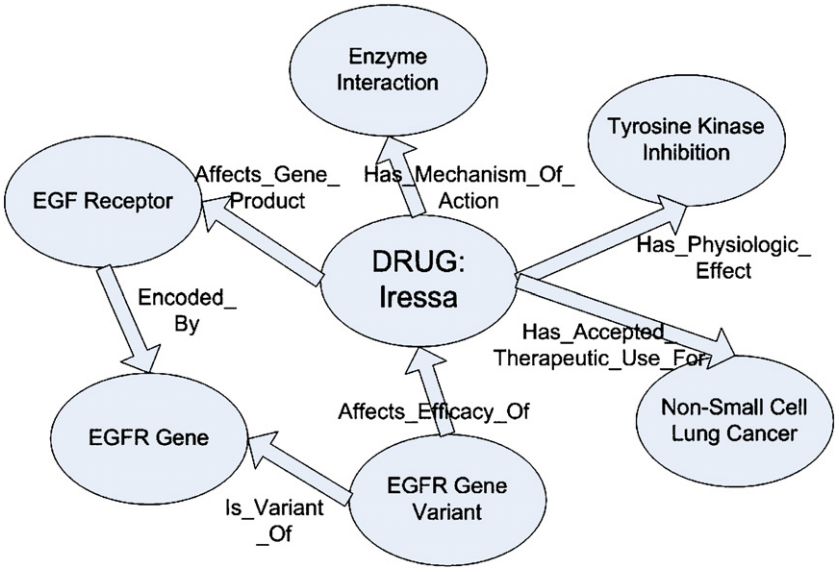

Fig. 1. Portion of semantic model relevant to pharmacokinetics and pharmacogenomics research.

Table 6
Genetic variants models and semantics [60–62]

Gene variant	A form of gene	Associated phenotype	Metabolic ratio	Variant frequency in population
CYP2C9*1	CYP2C9 gene	Affects efficacy of angiotensin II blockers	0.5	1–3% of Caucasian
CYP2D6*4	CYP2D6 gene	Affects efficacy of beta blockers	0	7% of Black 20% of Caucasian
CYP2D6*5	CYP2D6 gene	Affects efficacy of beta blockers	0	3% of Caucasian 5% of Black
CYP2D6*10	CYP2D6 gene	Affects efficacy of beta blockers	0	70% of Asian 8% of Caucasian
BRCA1 185delAG	Breast cancer 1 gene (BRCA1)	Hereditary breast ovarian cancer		3% of Ashkenazi Jewish breast cancer women
BRCA1 5382insC	Breast cancer 1 gene (BRCA1)	Hereditary breast ovarian cancer		0.75% of Ashkenazi Jewish breast cancer women
BRCA2 6174delT	Breast cancer 2 gene (BRCA1)	Hereditary breast ovarian cancer		3% of Ashkenazi Jewish breast cancer women
APC T3920A	Adenomatosis polyposis coli gene (APC)	Familial colorectal cancer		6% of Ashkenazi Jews
EGFR 2240dell8	Epidermal growth factor receptor gene	Affects efficacy of Iressa		

in significantly decreased drug metabolism, and Asian patients are often given lower doses of neuroleptic medications. Understanding the specific isoforms involved in drug metabolism should help physicians avoid potentially harmful drug interactions. Better characterization of drug metabolic profiles will allow for the prediction of potentially problematic drug interactions.

With advances in sequencing technology, allelic variations—such as single-nucleotide polymorphisms (SNPs) that have phenotypic consequences in patients—can be easily identified. The semantic relationships and description logic features of NCI Thesaurus provide a scaleable, multiaxial classification of concepts such as genetic variations, with semantic links between them and other hierarchies (e.g., therapeutic agents, diseases, and genes), supporting information representation and retrieval across all of these domains.

3.3. Pathways, genes, and gene products

Research into gene-interaction pathways for metabolic and regulatory functions is another rapidly growing area of research in cancer. KEGG and BioCarta are two well-known maintainers of this pathway information. The NCI Center for Bioinformatics recreates these pathways with live links to the genetic information in other databases as a feature of the Cancer Genome Anatomy project [46]. In the NCI Thesaurus, we include the pathway concepts and links to the genes and gene products involved in the pathways, where possible. Fig. 2 shows a portion of the NCI semantic model that includes these pathways concepts. Role relationships support retrieving information on a pathway of interest, identifying a gene of interest in that pathway, and finding any cancers known to be related to that gene or its products. Exploration of the pathways involved in a disease is supported by role relationships linking the disease to the gene or gene product hierarchy, and from there to the pathway hierarchy, perhaps going on to investigate the pathway in more detail at the CGAP website.

Similarly, NCI has a use case for linking alleles to diseases and polymorphisms to diseases for the NCICB supported Rembrandt project (REpository for Molecular BRAin Neoplasia DaTa), an informatics effort that is aimed at producing a national molecular/clinical database fully open and accessible to all investigators. NCICB is designing a robust bioinformatics knowledgebase framework, called the caIntegrator, that leverages data warehousing technology to host and integrate the clinical and functional genomics data. The NCI Thesaurus is part of the infrastructure that supports Rembrandt and caIntegrator. Creation of the allele terminology and linkage to diseases is in process.

3.4. Anatomy in humans and mice

Human diseases are frequently studied through the use of animal models. In the development of mouse models of human cancers and of other diseases, scientists need to annotate the information about the models, or slides and other research entities, with anatomic information. Mouse anatomy and human anatomy are somewhat different. For instance, mouse lungs have three lobes and human lungs have two lobes. Similarly, the mouse prostate and the human prostate are different. Some similar parts can have different functions in mouse and human. In order to facilitate comparative science, such comparative anatomy needs to be represented in the NCI Thesaurus for use by applications.

The gene expression database (GXD) has developed an ontology for adult mouse anatomy [47]. GXD is part of the larger Mouse Genome Informatics (MGI) Resource that provides the research community with integrated access to genetic, expression, and phenotypic data from the laboratory mouse. The GDX ontology will be used to describe

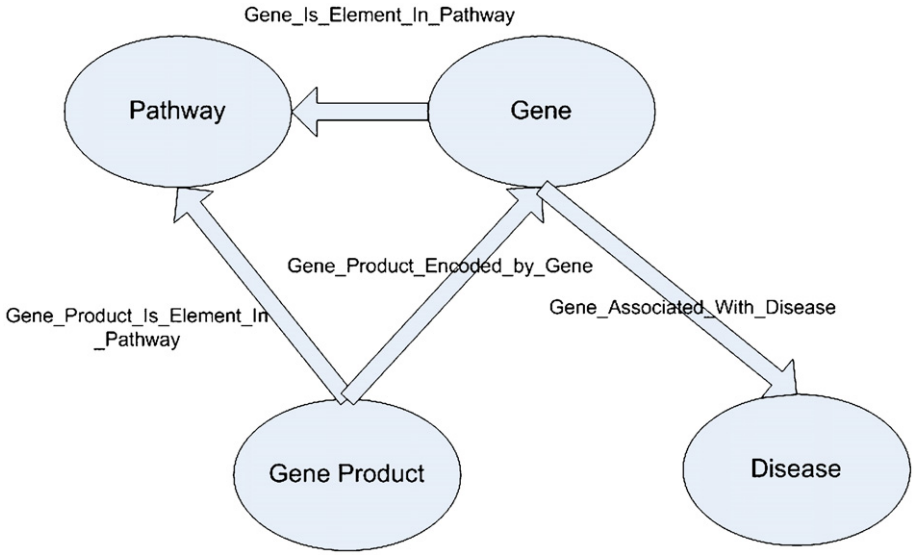

Fig. 2. Portion of semantic model relevant to metabolic and biochemical pathways research.

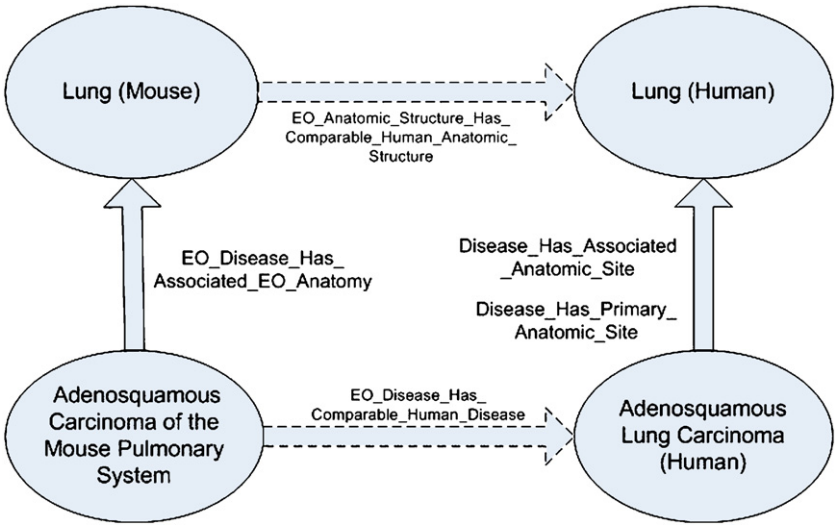

Fig. 3. Portion of semantic model relevant to mouse and human comparative science.

different types of data pertinent to mouse anatomy, including gene expression, biological process, phenotype, and pathology data, in a standardized and integrated fashion. The adult mouse anatomy is being included in NCI Thesaurus, and GXD is helping NCI map the human and mouse anatomies through a set of role relationships. This work will lead to a much closer integration of human and mouse cancer data, and thereby foster the study of human cancer and other diseases.

Fig. 3 shows a portion of the NCI Thesaurus semantic model as it would apply to the comparative anatomy and pathologic diagnoses of mice and humans for lung and adenosquamous carcinoma of lung. This model is not fully instantiated. The semantic links between mouse and human anatomy will have to be established at the lowest levels possible in the hierarchies, because the structures and functions of mice and humans are not identical, and the vocabulary structures are not identical. Hence, since role relationships inherit in Ontylog DL, role relationships asserted at a high level might incorrectly be inherited by child anatomic structure concepts to which the relationships do not apply. In the future, associations may be applied rather than roles, as these named relationships between concepts do not inherit. End users would be able to traverse the hierarchies and relationships in the same way they can with role relationships.

Similarly, the links between mouse diagnoses and human diagnoses are not instantiated yet. This will take careful scientific analysis of what links may appropriately be made between mouse and human diseases. Since mouse models are intentionally designed to model human diseases, the links will provide an important resource for cancer researchers.

4. Discussion

Many of the strengths and weaknesses of NCI Thesaurus directly reflect its nature and circumstances, and raise issues faced by other such efforts.

NCI Thesaurus arose as part of EVS Project efforts to satisfy NCI's controlled terminology needs, and this remains a core function. Apart from some areas covered adequately by other standard terminologies, NCI Thesaurus must cover the gamut of controlled terminology needed for widely varied research, clinical, administrative, and public information purposes. Often, 24-hour turnaround is required to meet the needs of dependent applications. While core content areas can be carefully planned and structured, and editing and technical procedures can provide reliability, there is inevitably some unevenness and inconsistency, especially in subject areas not viewed as mission-critical. At the same time, working closely with users, being driven by specific use cases, and publishing early and often, allows for the rapid feedback, correction of errors, and responsive growth that have been essential to success. Improved mechanisms for user feedback and evaluation will be important to keep this process on track.

Integration with the caCORE biomedical informatics infrastructure and the requirements of other users has pushed NCI Thesaurus increasingly into the realm of ontological knowledge representation. This was not an initial goal, and is emerging very unevenly from the minimal feature sets used for controlled terminology purposes. Only for types of cancer is a systematic effort underway to represent a full set of biomedically defining features. Molecular concepts and relationships are heavily represented, but only selective subsets relevant to cancer, with links to authoritative external sources such as GenBank [48], OMIM [49], and UniProt [50]. How to better integrate with such external resources remains a great, largely unsolved issue.

Software choices have also been important, especially in shaping the ontological aspects of NCI Thesaurus. The Ontylog DL and tools met the initial needs of NCI Thesaurus development and were the only production quality tools available then for terminology development, especially in providing a terminology server and workflow support

for distributed editing. Ontylog DL intentionally sacrificed more powerful knowledge representation for scalability. Such considerations led to their use in similar efforts at SNOMED [8,51,52], Kaiser Permanente [53], and the Department of Veteran's Affairs for its National Drug File Reference Terminology (NDF-RT) [54,55]. Other development environments could support more powerful knowledge representation—Galen used GRAIL and, later, OILED tools [56,57] and the Foundational Model of Anatomy used frame representation in Protégé [58,59]. The new Protégé OWL plug-in can be used in conjunction with classifiers such as RACER or FaCT that implement highly expressive description logics. Our project is actively pilot testing Protégé/Racer for example. To date, none of these tools have met the technical requirements needed to support the multi-editor, distributed development model used to maintain and extend NCI Thesaurus. In addition to terminology development, NCI requires terminology server resources. An advantage of Ontylog DL has been the availability of a commercial, off-the-shelf terminology server for it.

Nonetheless, reliance on Ontylog DL and associated software products has involved significant software limits and design compromises. Some information, such as numeric or other specifications relevant to many of the *Disease_May_Have* roles, has so far been omitted. Logical negation has been expressed through *Excludes* roles, separate name concepts (e.g., Absence of Stromal Invasion), or else lost. Role groups are (temporarily) being used to label prognostic subgroups with distinctive molecular traits, where correctly these should be split out into separate sets or concepts. None of this has prevented satisfying current use-cases or correctly classifying the terminology, and some aspects can be corrected in exporting NCI Thesaurus data to more expressive forms such as OWL/DL.

Open, publicly available production quality terminology development and terminology server tools that support expressive description logic are needed. Open tools and servers seem to be critical to biomedicine moving to open, collaborative terminology development. While there are many unresolved issues associated with collaborative terminology development, the expense of terminology development and the range of subject matter expertise that terminology development requires, places it beyond the capability of even large organizations. Some way to share the burden of terminology creation and maintenance is going to be needed, as is a way to capitalize on expertise of contributors working across institutional boundaries.

5. Conclusion

Cancer research and clinical practice increasingly require tight integration of large amounts of molecular and clinical data. The NCI Thesaurus has been extended to support such integration, most notably in its logic-based characterization of types of cancer, but also—as outlined above—in such areas as drugs, molecular pathways, and the comparative anatomy of experimental organisms. Translational research on cancer now requires this sort of explicit knowledge representation, and it is being built into the biomedical informatics infrastructure on which that research increasingly depends. Formal use cases help decide the priorities and design for such knowledge representations, which are far from formal completeness but designed to meet the information needs of the cancer community. This combination of controlled terminology and ontology has so far proved a useful, if impure, hybrid.

Acknowledgments

The authors thank their colleagues on the EVS project, Gilberto Fragoso, Jim Oberthaler, Nicole Thomas, Paula Fry, and the other domain expert editors, and Subha Madhavan, manager of the NCICB Rembrandt project as well as Martin Ringwald and Terry Hayamizu of the Mouse Genome Informatics Resource at Jackson Laboratories. We also acknowledge the support of Ken Buetow and Peter Covitz at the NCICB, and of Gisele Sarosy and Richard Manrow at the NCI Office of Communications.

References

[1] Hartel FW, De Coronado S. Information standards within the National Cancer Institute. In: Silva JS, Ball MJ, Chute CG, Douglas JV, Langlotz C, Niland J, Scherlis W, editors. Cancer informatics: essential technologies for clinical trials. Berlin: Springer; 2002. p. 135–56.

[2] Hubbard SL. Information systems in oncology. In: Devita VT, Hellman S, Rosenberg SA, editors. Cancer: Principles & Practice of Oncology. 6th ed. Philadelphia: Lippincott Williams & Wilkins; 2001. p. 3135–46.

[3] caBIG. Available at: <http://cabig.nci.nih.gov/>. Accessed October 14, 2004.

[4] Sittig DF. Grand challenges in medical informatics. J Am Med Inform Assoc 1994;1(5):412–3.

[5] Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. Meth Inf Med 1993;32:281–91.

[6] U.S. Department of Health & Human Services. Federal Health Architecture. <http://www.hhs.gov/fedhealtharch/>. Accessed October 14, 2004.

[7] Rector A, Nowlan W, Glowinski A. Goals for concept representation in the GALEN project. In: Proc Annu Symp Comput Appl Med Care 1993:414–18.

[8] SNOMED International, College of American Pathologists. SNOMED CT. Available from: <http://www.snomed.org/snomedct/index.html/>. Accessed February 23, 2005.

[9] Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. Nat Rev Genet 2004;5(3):213–22.

[10] Covitz P, Hartel FW, Schaefer C, de Coronado S, Fragoso G, Sahni H, et al. caCORE: A common infrastructure for cancer informatics. Bioinformatics 2003;19:2402–12.

[11] De Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. Medinfo 2004;2004:33–7.

[12] Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. J Biomed Inform 2005;38(2):114–29.

[13] Hartel FW, Fragoso G, Ong K, Dionne R. Enhancing quality of retrieval through concept edit history. In: Musen M, editor. AMIA Annu Symp Proc 2003:279–83.

[14] The NCICB Core Infrastructure Group. Available at: <http://ncicb.nci.nih.gov/core/>. Accessed October 14, 2004.
[15] Welcome to the NCI Terminology Browser. Available at: <http://nciterms.nci.nih.gov/>. Accessed October 14, 2004.
[16] Available at: <ftp://ftp1.nci.nih.gov/pub/cacore/EVS/>. Accessed October 14, 2004.
[17] Golbeck J, Fragoso G, Hartel FW, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. J Web Semantics 2003;1:75–80.
[18] Available at: <ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/>. Accessed October 14, 2004.
[19] Maes B, De Wolf-Peeters C. Marginal zone cell lymphoma: an update on recent advances. Histopathology 2002;40(2):117–26.
[20] Starostik P, Patzner J, Greiner A, Schwarz S, Kalla J, Ott G, et al. Gastric marginal zone B-cell lymphomas of MALT type develop along 2 distinct pathogenetic pathways. Blood 2002;99(1):3–9.
[21] Dierlamm J, Baens M, Wlodarska I, Stefanova-Ouzounova M, Hernandez JM, Hossfeld DK, et al. The apoptosis inhibitor gene AP12 and a novel 18q gene, MLT, are recurrently rearranged in the t(11;18)(q21;q21) associated with mucosa-associated lymphoid tissue lymphomas. Blood 1999;93(11):3601–9.
[22] Liu H, Ruskon-Fourmestraux A, Lavergne-Slove A, Ye H, Molina T, Bouhnik Y, et al. Resistance of t(11;18) positive gastric mucosa-associated lymphoid tissue lymphoma to *Helicobacter pylori* eradication therapy. Lancet 2001;357(9249):39–40.
[23] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identifies by gene expression profiling. Nature 2000;403(6769):503–11.
[24] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Med 2002;8(1):68–74.
[25] Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. N Engl J Med 2002;346(25):1937–47.
[26] McLean TW, Ringold S, Neuberg D, Stegmaier K, Tantravahi R, Ritz J, et al. TEL/AML-1 dimerizes and is associated with a favorable outcome in childhood acute lymphoblastic leukemia. Blood 1996;88(11):4252–8.
[27] Borkhardt A, Cazzaniga G, Viehmann S, Valsecchi MG, Ludwig WD, Burci L, et al. Incidence and clinical relevance of TEL/AML1 fusion genes in children with acute lymphoblastic leukemia enrolled in the German and Italian multicenter therapy trials. Blood 1997;90(2):571–7.
[28] Uckun FM, Pallisgaard N, Hokland P, Navara C, Narla R, Gaynon PS, et al. Expression of TEL-AML1 fusion transcripts and response to induction therapy in standard risk acute lymphoblastic leukemia. Leuk Lymphoma 2001;42(1–2):41–56.
[29] Kanerva J, Saarinen-Pihkala UM, Niini T, Riikonen P, Mottonen M, Makipernaa A, et al. Favorable outcome in 20-year follow-up of children with very-low-risk ALL and minimal standard therapy, with special reference to TEL-AML1 fusion. Pediatr Blood Cancer 2004;42(1):30–5.
[30] Pui CH, Evans WE. Acute lymphoblastic leukemia. N Engl J Med 1998;339(9):605–15.
[31] Heerema NA, Sather HN, Sensel MG, Zhang T, Hutchinson RJ, Nachman JB, et al. Prognostic impact of trisomies of chromosomes 10, 17, and 5 among children with acute lymphoblastic leukemia and high hyperdiploidy (>50 chromosomes). J Clin Oncol 2000;18(9):1876–87.
[32] Arico M, Valsecchi MG, Camitta B, Schrappe M, Chessells J, Baruchel A, et al. Outcome of treatment in children with Philadelphia chromosome-positive acute lymphoblastic leukemia. N Engl J Med 2000;342(14):998–1006.
[33] Schrappe M, Arico M, Harbott J, Biondi A, Zimmermann M, Conter V, et al. Philadelphia chromosome-positive (Ph+) childhood acute lymphoblastic leukemia: good initial steroid response allows early prediction of a favorable treatment outcome. Blood 1998;92(8):2730–41.
[34] Ribeiro RC, Broniscer A, Rivera GK, Hancock ML, Raimondi SC, Sandlund JT, et al. Philadelphia chromosome-positive acute lymphoblastic leukemia in children: durable responses to chemotherapy associated with low initial white blood cell counts. Leukemia 1997;11(9):1493–6.
[35] Rubnitz JE, Look AT. Molecular genetics of childhood leukemias. J Pediatr Hematol Oncol 1998;20(1):1–11.
[36] Pui CH, Frankel LS, Carroll AJ, Raimondi SC, Shuster JJ, Head DR, et al. Clinical characteristics and treatment outcome of childhood acute lymphoblastic leukemia with the t(4;11)(q21;q23): a collaborative study of 40 cases. Blood 1991;77(3):440–7.
[37] Crist WM, Carroll AJ, Shuster JJ, Behm FG, Whitehead M, Vietti TJ, et al. Poor prognosis of children with pre-B acute lymphoblastic leukemia is associated with the t (1;19) (q23;p13): a Pediatric Oncology Group study. Blood 1990;76(1):117–22.
[38] Hunger SP. Chromosomal translocations involving the E2A gene in acute lymphoblastic leukemia: clinical features and molecular pathogenesis. Blood 1996;87(4):1211–24.
[39] Fletcher CD, Unni KK, Mertens F, editors. Pathology and genetics of tumours of soft tissue and bone. Lyon, France: IARC Press; 2002.
[40] Kleihues P, Cavenee WK, editors. Pathology and Genetics of Tumours of the Nervous System. Lyon, France: IARC Press; 2000.
[41] Vitali R, Cesi V, Nicotra MR, McDowell HP, Donfrancesco A, Mannarino O, et al. c-Kit is preferentially expressed in MYCN-amplified neuroblastoma and its effect on cell proliferation is inhibited in vitro by STI-571. Int J Cancer 2003;106(2):147–52.
[42] Ricotti E, Fagioli F, Garelli E, Linari C, Crescenzio N, Horenstein AL, et al. c-kit is expressed in soft tissue sarcoma of neuroectodermic origin and its ligand prevents apoptosis of neoplastic cells. Blood 1998;91(7):2397–405.
[43] Chute CG, Carter JS, Tuttle MS, Haber MW, Brown SH. Integrating pharmacokinetics knowledge into a drug ontology as an extension to support pharmacogenomics. Proc AMIA Symp 2003:170–74.
[44] Hirsch ER, Witta S. Biomarkers for prediction of sensitivity to EGFR inhibitors in non-small cell lung cancer. Curr Opin Oncol 2005;17(2):118–22.
[45] Amador ML, Oppenheimer D, Perea S, Maitra A, Cusati G, Iacobuzio-Donahue C, et al. An epidermal growth factor receptor intron 1 polymorphism mediates response to epidermal growth factor receptor inhibitors. Cancer Res 2004;64(24):9139–43.
[46] Pathways. Available at: <http://cgap.nci.nih.gov/Pathways/>. Accessed October 14, 2004.
[47] Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. Genome Biol 2005;6(3):R29.
[48] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2005;1:33. [Database Issue: D34-8].
[49] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005;1:33. [Database Issue: D514-7].
[50] Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;1:33. Database Issue:D154-9.
[51] Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. Proc AMIA Annu Fall Symp 1997:640–44.
[52] Spackman KA, Campbell KE. Compositional concept representation using SNOMED: toward further convergence of clinical terminologies. Proc AMIA Symp 1998:740–44.
[53] Dolin RH, Mattison JE, Cohn S, Campbell KE, Wiesenthal AM, Hochhalter B, et al. Kaiser Permanente's convergent medical terminology. Medinfo 2004;2004:346–50.
[54] Carter JS, Brown SH, Erlbaum MS, Gregg W, Elkin PL, Speroff T, et al. Initializing the VA medication reference terminology using UMLS Metathesaurus co-occurrences. Proc AMIA Symp 2002: 116–20.

[55] Brown SH, Elkin PL, Rosenbloom ST, Husse r C, Bauer BA, Lincoln MJ, et al. VA national drug file reference terminology: a cross-institutional content coverage Study. Medinfo 2004;2004:477–81.
[56] Rector AL, Rogers JE. Ontological issues in using a description logic to represent medical concepts: experience from GALEN: Part 1—Principles. IMIA WG6 Workshop: Phoenix, Arizona, Nov. 1999. Available at: <http://www.opengalen.org/info/IMIAWG6-1999.pdf/>.
[57] Zanstra PE, van der Haring EJ, Flier F, Rogers JE, Solomon WD. Using the GRAIL language for classification management. Stud Health Technol Inform 1997;43(Pt. A):441–5.
[58] Rosse C, Mejino Jr JL. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36(6):478–500.
[59] Michael J, Mejino Jr JL, Rosse C. The role of definitions in biomedical concept representation. Proc AMIA Symp. 2001; 463–67.
[60] Fodor FH, Weston A, Bleiweiss IJ, McCurdy LD, Walsh MM, Tartter PI, Brower ST, Eng CM. Frequency and carrier risk associated with common BRCA1 and BRCA2 mutations in Ashkenazi Jewish breast cancer patients. Am J Hum Genet 1998;63:45–51.
[61] Gaedigk A, Ndjountche L, Gaedigk R, Leeder JS, Bradford LD. Discovery of a novel nonfunctional cytochrome P450 2D6 allele, CYP2D642, in African American subjects. Clin Pharmacol Ther 2003;73(6):575–6.
[62] Wan Y-J, Poland RE, Han G, Konishi T, Zheng Y-P, Lin K-M. Analysis of the CYP2D6 gene polymorphism and enzyme activity in Africa-Americans in Southern California. Pharmacogenetics 2001;11(6):489–99.