July 23, 2004

National Cancer Institute
External Data Standards Review

Prepared for:

National Cancer Institute Center for Bioinformatics
National Institutes of Health
6116 Executive Boulevard, Suite 403
Rockville, MD 20852

Technical Project Officer:

Peter Covitz, Ph.D.
Director, Bioinformatics Core Infrastructure
NCI Center for Bioinformatics
covitzp@mail.nih.gov

Prepared by:

Science Applications International Corporation
6565 Arlington Boulevard
Falls Church, VA 22042

CONTENTS

EXHIBITS

EXTERNAL DATA STANDARDS REVIEW

1.0 INTRODUCTION

The plan to harmonize the Common Data Elements (CDEs) in the National Cancer Institute (NCI) Cancer Data Standards Repository (caDSR) as well as the controlled terminology curated, developed, and maintained through the NCI's Enterprise Vocabulary Services (EVS) involves assessing national and international standards for use in data collection instruments and related databases. This report, which has been produced as part of the implementation of the Tactical Action Plan for CDE Harmonization, reviews potentially applicable data and vocabulary standards that are, for the most part, developed or defined by organizations outside of the NCI. The report begins with a brief description of the harmonization project, describes the role of external standards in NCI data management, and explains how the standards have been selected, reviewed, and categorized. The report also includes a description of some of the standards review and approval bodies whose recommendations are pertinent to the NCI. The report also includes a few standards developed by NCI programs that are used by a number of organizations across the cancer research community. Each standard is described with a brief summary, and other pertinent information such as the sponsor, version information, and a description of its applicability.

The list of potentially applicable standards is based on a review of the data elements and data concepts currently registered in the caDSR or included in EVS as well as those recommended by NCI personnel, including caDSR Context Administrators and EVS staff. The report includes recommendations for how NCI should engage in external standards efforts, as well as which standards should be considered by NCI for adoption, or use in informing CDE development and the use of controlled terminologies. The potential for registration in the caDSR or inclusion in EVS is also addressed.

1.1 Background

The NCI is supporting a broad initiative to develop standard tools and practices that include controlled vocabularies, CDEs, and logical models of entities within and across life science domains. The use of common information building blocks for data capture and reporting facilitates the understanding and sharing of cancer research information. To support these initiatives, the NCI Center for Bioinformatics (CB) has developed a core infrastructure (caCORE) consisting of EVS, jointly maintained with the NCI Office of Communications, in order to support the development of well-defined terminologies; the Cancer Bioinformatics Infrastructure Object (caBIO), a use-case and model-driven architecture for biomedical data management; and the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 11179-based caDSR to record and manage contextual metadata for CDEs.

The caDSR and EVS were created to improve the ability to share and exchange data to advance basic and clinical cancer research by standardizing CDEs and terminology used in data capture and reporting of cancer clinical trials and other basic research. They were designed to overcome problems that prevent comparative analysis such as incompatibility in the naming and contents of clinical trials report forms and inconsistent documentation of data in different repositories.

The goal of the caDSR in particular is to define a comprehensive set of standardized metadata descriptors for cancer research terminology and clinical trial protocols and forms. Various NCI offices in conjunction with sponsored clinical trial organizations have developed the content of the caDSR. CDEs have been developed in domain-specific work areas referred to as "Contexts."

A "harmonization team" was created and tasked with proposing a process for harmonization, curation, and governance of caDSR content. Part of the harmonization initiative is reviewing standards developed outside NCI that may be applicable to NCI information. The team has reviewed health information standards and other standards commonly used in sharing and reporting information. The report reviews their applicability to NCI data, and makes recommendations for adoption, as well as for involvement of NCI in further development of the standards.

1.2 Role of External Standards

Organizations use external standards to save time and resources that might be spent in creating and maintaining long lists of values. The use of the standards also ensures a common understanding of data being exchanged. A popular example of an external standard is the Federal Information Processing Standard (FIPS) 10-4, a list of Countries, Dependencies, Areas of Special Sovereignty, and their Principal Administrative Divisions, that maintains information on the basic geopolitical entities in the world, together with the principal divisions that comprise each entity. It is an example of an authoritative source for specific information that is commonly used, which means that every organization does not have to keep current with changes to the information. Other lists commonly used over time have included metropolitan statistical areas, Zone Improvement Plan (ZIP) code lists, standard units of measure, race and ethnicity identifiers, and the like.

Data exchange standards (also called transaction standards or messaging standards) provide specifications for format and content of data exchanges. These standards facilitate data sharing through the application of a commonly understood and accepted template. They specify format and meaning of data elements involved in data exchanges. Through use, organizations reduce costs spent on data acquisition and conversion due to a higher level of interoperability with data suppliers.

The Committee on Health Data Standards of the Data Council of the Department of Health and Human Services (HHS) states:

"One of the biggest issues for health data today is the lack of shared data standards. The lack of shared standards increases paperwork and data collection burdens, and reduces the analytic potential of health data. Without consistent use of data standards, there is little ability to make multiple uses of or link data, which limits the usefulness of the HHS data to our public and private data customers and State partners, and vice versa. The need for shared health data standards encompasses the need for better agreement on common health data vocabularies, assurances of privacy, and other issues surrounding electronic transmission of information." <http://aspe.hhs.gov/datacncl/hdscmte.htm>

With the need to improve information interchange in the health arena, a number of standards have been developed to standardize the way organizations record disease types, health care provider information, and patient information. The common adoption of these standards will help to ensure that information can be seamlessly exchanged, and that data can be reused. For example use of standard data coding schemes in clinical trials will make it possible to combine data from a number of trials in cross-study analyses. Use of common data formats makes it possible to develop programming interfaces for the automated exchange of clinical or bioinformatic information between related databases.

Also driving the adoption of data standards is the Consolidated Health Informatics (CHI) initiative that aims to "establish a portfolio of existing clinical vocabularies and messaging standards enabling federal agencies to build interoperable federal health data systems." CHI standards will work in conjunction with the Health Insurance Portability and Accountability Act (HIPAA) transaction records and code sets and HIPAA security and privacy provisions. About 20 department/agencies including HHS, Department of Veterans Affairs (VA), Department of Defense (DOD), Social Security Administration (SSA), General

Services Administration (GSA), and the National Institute of Standards and Technology (NIST) are active in the CHI governance process. Through this process, all federal agencies will incorporate the adopted standards into their individual agency health data enterprise architecture used to build all new systems or modify existing ones. There are now over 20 CHI standards. Information on the current list of standards can be found on the CHI web site at (http://www.whitehouse.gov/omb/egov/gtob/health_informatics.htm/). On January 24, 2004, the National Committee on Vital and Health Statistics (NCVHS) reviewed a set of CHI recommendations (See (http://www.ncvhs.hhs.gov/040129lt.pdf/). It is anticipated that NCVHS will review each set of CHI recommendations. A new set of recommendations was published in May 2004.

NCI is obliged to use some standards in the collection, storage, and reporting of information. For example, the HHS has mandated the use of Health Level Seven (HL7) and the Logical Observation Identifiers Names and Codes (LOINC) standards. In time, HHS may adopt the other CHI standards. The Food and Drug Administration (FDA) is requiring the use of standards in reporting clinical trials data. The Clinical Data Interchange Standards Consortium (CDISC) and the HL7 Regulated Clinical Research Information Management (RCRIM) are harmonizing to reach consensus on this standard for representation of clinical trial protocol elements. The Office of Management and Budget (OMB) specifies the use of certain race and ethnicity standards in the reporting of information about people.

Appendix D presents some information about the standards review and approval boards, their spheres of influence, and the types of standards that they review for adoption or approval.

1.3 Categories of External Standards

Standards can be categorized in various ways. However, for the purposes of this report, the standards have been categorized into one of three types: 1) Common Demographic/Information Processing Standards and Code Sets, 2) Health-related Vocabulary/Coding Standards, and 3) Health-related Transaction Standards. The standards are presented in three primary groups based on relevance to NCI. Appendix A of the document describes those standards that are recommended for serious consideration by NCI for adoption and use. Appendix B describes those standards that should be considered by NCI, but which may not be sufficiently mature or may duplicate some of the standards in Appendix A. Appendix C includes those standards that were reviewed but found unlikely to meet NCI needs. Appendix D lists standards approval bodies that review and make recommendations about standards use.

Demographic/Information Processing standards and code sets are data standards that apply generally across federal agencies and may be of use to any organization conducting surveys and statistical analysis. These standards include demographic characteristics, locational information, formats for date/time and the like.

Vocabulary standards deal primarily with the standardization of terms and definitions for concepts in a particular business area. The health care arena has a host of ontologies and thesauri that have published terms and definitions. The NCI has sought to manage its vocabulary resources using EVS. EVS publishes and provides server access for individual vocabularies, brings together multiple vocabulary resources from various sources into a single collection through the NCI Metathesaurus, and produces the NCI Thesaurus, a description logic vocabulary which supplies NCI preferred terms, synonyms, definitions and associated relationships between standardized concepts in the NCI domain. By NCI Executive Group charter, EVS is expected to provide for all the terminology needs of NCI. Required code sets are to be provided through this central resource that manages licensing, maintenance and update issues for NCI, as well as centralizing required terminology sets for dependent applications.

Transaction (or data exchange) standards specify not only a concept and definition, but the physical attributes of the information, and usually include field length and type and possibly related permissible value lists. They support the standardization of data exchanges, providing clear documentation to support efficient and meaningful data exchange. The products of transaction standards include a) definition of the circumstances (trigger event) under which data is passed, b) definition of the data to be passed in a specific circumstance, and c) definition of the technical requirements for communication. In many cases, transaction standards include both data element metadata as well as lists of terms and definitions. As a result, their scope may overlap that of the terminology standards. Examples of health data element standards include HL7 and those from CDISC.

The report also describes a couple of standards that are developed and maintained by NCI programs for use across the cancer research community. These include the Common Terminology Criteria for Adverse Events (CTCAE), a grading system for reporting the acute and late effects of cancer treatment that was developed by the Cancer Therapy Evaluation Program (CTEP), as well as the NCI Thesaurus that is part of EVS. NCI has chosen to manage CDEs in the caDSR, and to make use of the terminology in EVS in the development of CDEs for use in form and application design. NCI is in the process of recording external standards for discovery and reuse in the caDSR. Therefore, together, the caDSR and associated controlled terminology in EVS, will be the source of NCI standards, which will be data elements and vocabularies preferred for use in case report form and system design. The caDSR will apply “registration statuses” to CDEs to identify candidate and approved NCI standards.

1.4 References

The following references were guidance for this document:

- *Address Data Content Standard*, Public Review Draft, Subcommittee on Cultural and Demographic Data, Federal Geographic Data Committee, April 17, 2003, Version 2.
- American National Standards Institute (ANSI) Accredited Standards Committee X12 (X12) <http://www.x12.org/x12org/index.cfm/>.
- American Society for Testing and Materials (ASTM) Standard Specification for Coded Values Used in the Electronic Health Record <http://www.astm.org/>.
- Annual Demographic Survey, March Supplement Data Dictionary, <http://www.bls.census.gov/cps/basic/datadict/199801/puf98dd.htm/>.
- Bureau of the Census Current Population Survey <http://www.bls.census.gov/cps/cpsmain.htm/>.
- *caDSR Business Rules*, August 22, 2003.
- *caDSR Designation Rules*, August 22, 2003.
- *caDSR Version Rules*, August 22, 2003.
- caCORE 1.0 caDSR Database API, National Cancer Institute Center for Bioinformatics, National Institutes of Health, U.S. Department of Health and Human Services, 2002.
- *CaCORE 1.0 Technical Guide*. National Cancer Institute Center for Bioinformatics, National Institutes of Health, U.S. Department of Health and Human Services, 2002.
- caCORE Project Plans 2003-2004, NCI Center for Bioinformatics, June 5, 2003.
- Cancer.gov, *Development of a Cancer Biomedical Informatics Grid (caBIG)*, <http://cabig.nci.nih.gov/overview/>.
- CDC National Center for Health Statistics (NCHS) <http://www.cdc.gov/nchs/>.
- CDC National Electronic Telecommunications System for Surveillance (NETSS) <http://www.cdc.gov/epo/dphsi/netss.htm/>.
- Centers for Disease Control (CDC) National Committee on Vital and Health Statistics (NCVHS) <http://ncvhs.hhs.gov/>.

- Clinical Data Interchange Standards Consortium (CDISC) <http://www.cdisc.org/standards/index.html>.
- *Common Data Elements Implementation Guide, Version 2.4*, Standards and Liaison Committee Health Information and Surveillance Systems Board, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
- Consolidated Health Informatics Initiative (CHI) <http://www.whitehouse.gov/omb/egov/gtob/health_informatics.htm/>.
- Current Procedural Terminology (CPT)-4. <http://www.ama-assn.org/ama/pub/category/3113.html/>.
- Food and Drug Administration (FDA) Center for Drug Evaluation and Research (CDER) Data Standards Manual <http://www.fda.gov/cder/dsm/>.
- Gillman, Daniel W. and Martin V. Appel. "The Statistical Metadata Repository: an electronic catalog of survey descriptions at the U.S. Census Bureau," *IASSISST Quarterly*, summer 1997.
- Havener L, Hultstrom D, editors. Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Ninth Edition, Version 10.2. Springfield Illinois: North American Association of Central Cancer Registries, March 2004.
- Health Level Seven (HL 7) <http://www.h17.org/>.
- International Organization for Standardization (ISO). <http://www.iso.org/>.
- ISO/IEC 11179-3, Information technology -- Metadata Registries (MDR) -- Part 3: Registry Metamodel and basic attributes, Second edition, February 2003.
- ISO/IEC TR 20943-1, Technical Report, Information technology -- Procedures for achieving metadata registry (MDR) content consistency -- Part 1: Data elements, First edition, August 2003.
- ISO/IEC DTR 20943-3, Draft Technical Report, Information technology -- Procedures for achieving metadata registry content consistency -- Part 3: Value domains, June 2003.
- National Cancer Institute (NCI) Cancer Therapy Evaluation Program (CTEP), Common Terminology Criteria for Adverse Events v 3.0 (CTCAE), December 12, 2003.
- National Committee on Vital and Health Statistics, January 29, 2004, *Consolidated Health Informatics Initiative, Final Recommendations.*
- National Council on Vital and Health Statistics (NCVHS) <www.ncvhs.hhs.gov/>.
- National Institute of Standards and Technology (NIST) Federal Information Processing Standards (FIPS) <http://www.itl.nist.gov/fipspubs/>.
- Office of Management and Budget (OMB) Standards for the Classification of Federal Data on Race and Ethnicity <http://www.whitehouse.gov/omb/fedreg/ombdir15.html/>.
- Public Health Information Network (PHIN) <http://www.cdc.gov/phin/>.
- Release Notes, caCORE Version 1.2, June 13, 2003, National Cancer Institute Center for Bioinformatics.
- Request for Quotation 72663, Statement of Work, Common Data Element Development and Harmonization for National Cancer Institute Center for Bioinformatics, May 6, 2003.
- Standard Occupational Classification (SOC) System <http://www.bls.gov/soc/>.
- Taylor, Barry N., 1995: *Guide for the Use of the International System of Units (SI)*, NIST Special Publication No. 811, (Supersedes 1991 Edition), National Institute of Standards and Technology (NIST) (NIST SP 811 PDF version)
- U.S. Food and Drug Administration, Center for Drug Evaluation and Research (CDER) <http://www.fda.gov/cder/dsm/>.
- United States Postal Service (USPS) publication 28, *Postal Addressing Standards,* <http://pe.usps.gov/cpim/ftp/pubs/Pub28/pub28.pdf/>.

1.5 Report Contributors

This report was prepared for the NCI by Science Applications International Corporation (SAIC) as part of the CDE Development and Harmonization contract. Kathleen Gundry was the principal author, and team contributors included: Tommie Curtis (SAIC), Dr. Hong Dang (Alpha Gamma Technologies), Jocelyn Leatherwood (SAIC), Mead Walker (consultant), and Dr. Edward "Tin" Wong (SAIC). Other major reviewers and contributors included Margaret Haber of NCI's Office of Communications, Frank Hartel (NCICB), and Peter Covitz (NCICB).

2.0 METHODOLOGY

An initial list of potentially applicable data standards was developed through research into health data standards and from suggestions made by caDSR Context Administrators and EVS staff who have experience with data standards in their particular research domains. Access to some standards was gained through Frank Hartel (hartelf@mail.nih.gov), NCI's external standards coordinator. The report was assembled from individual reports from a number of team members who researched the current status of the standards and assessed their applicability to NCI.

3.0 RECOMMENDATION SUMMARY

This section includes recommendations on using external standards in developing NCI common data elements, a process for registering and maintaining standards metadata in the caDSR and adding and maintaining vocabulary in the EVS to provide semantic consistency to the metadata. It also contains suggestions for NCI involvement in external standards organizations. It also includes some analysis of the standard content, identifying overlaps and conflicts.

NCI has a mandate to comply where feasible with standards from CHI and NCVHS, which is sponsored by HHS, as the department's statutory [42 U.S.C. 242k(k)] public advisory body on health data, statistics and national health information policy (as informed by CHI). The NCVHS actively reviews and evaluates other external standards for applicability within HHS, and has adopted standards such as HL7. These standards approval bodies also recommend other federal standards such as those of Centers for Disease Control (CDC), FDA, VA, and others, and NCI will seriously consider those for adoption.

Recommendations for categorization of the standards were based on the following principles:

- NCI will adopt standards that facilitate standardized data exchange with its information trading partners.
- NCI will adopt appropriate public, open access or nationally licensed data standards [that is, that are freely available and do not involve license fees for NCI or its information trading partners] where they are available.
- NCI will adopt standards that have been tested and are used widely in the relevant communities.

3.1 Recommendation Context

Perhaps you've heard that "The nice thing about standards is that there are so many to choose from" (Andrew S. Tannenbaum). Inherent in the quip is the recognition that having multiple standard ways to do something can be almost as hard to manage as having no standard way at all. NCI will devote considerable resources towards making standards available to researchers, and participants. It is important that NCI: a) make clear how the individual standards are going to be used, and b) make particular standards available for particular purposes.

Standards for data and for vocabulary will be used to bring consistency to data that is submitted to the NCI. The selection of standards should be based on 3 criteria: 1) consistency with the data representations used in the scientific setting, 2) ease of implementation for researchers and other providers of data to NCI, and 3) support of back end uses of the data at NCI. In general, this leads to the requirement that the chosen standards support the breadth of NCI data requirements, be consistent or the same as those generally used in clinical and research settings, and be easy to acquire and use.

3.2 Standards Coordination

NCI should endeavor to have a presence in those bodies that recommend the use of standards in order to ensure that chosen standards have the desired breadth and functionality.

3.3 Standards Registration and Approval

The caDSR and EVS should become the authoritative sources within NCI for accessing recommended data standards. Initially, various data standards should be registered in the caDSR to make them available for designation by Contexts, and for use in clinical form design and database development. The data standards will be subject to the harmonization process, whereby Contexts will review them for adoption as NCI standards. Reviewing Contexts will need to consider the recommendations of the CHI and the NCVHS for HHS. All standards in Appendix A are considered to be relevant to NCI, and are recommended for registration in the caDSR and for consideration as NCI standards. The NCI can subsequently take up review of the standards in Appendix B.

Sets of vocabulary and code sets will be maintained, mapped and developed in EVS. They will be used as a primary source for development of data elements in the caDSR, and may serve as referenced value domains.

As each standard is reviewed for approval for use across NCI, a decision will need to be made about the most efficient and effective means of managing the standard content. Currently, all vocabulary collections and many code sets are managed in the EVS. Those terms are available to the caDSR for use in forming metadata components. A code set that is also a permissible value set for a data element registered in the caDSR can be registered in the caDSR as a value domain. Another option is to reference an external source where a list of terms or codes is stored. These decisions will be made on a case-by-case basis as standards are reviewed, approved, and implemented at the NCI.

3.4 Standards Categories

Exhibit 1 summarizes the various standards categorized by type. The recommendations column indicates to which recommendation category the standard has been assigned. It is recommended that the harmonization be undertaken by analysis and comparison of standards within the identified categories. In some cases, more than one standard within a category will need to be approved for adoption, as the standards may be complementary, not competing. In the following sections, the team presents initial recommendations on standards within the categories, with A indicating recommended for NCI use, B indicating recommendation for further consideration by NCI, and C indicating that the standard is not recommended at this time. There are no recommendations for the standards review and approval bodies. NCI will seriously consider recommendations from all listed standards review organizations.

Exhibit 1. Standard Type Table

Standard Type	**Standard Name**	**Standard Content**	**Recommendation**[1]
Common Demographic/Information Processing and Code Sets			
Address	Federal Geographic Data Committee (FGDC) Address Data Content Standard (Draft)	Address information for a physical location	C
	FIPS 5-2, Codes for Identification of the States, District of Columbia, and Outlaying Areas of the United States, and Associated Areas	State names	A
	FIPS 10-4, Countries, Dependencies, Areas of Special Sovereignty, and their Principal Administrative Divisions	Country names	A
	ISO 3166-1, Country Codes	Country codes	A
	ISO 11180:1993, Postal Addressing	Address format	A
	Universal Postal Union	Address format, state codes, country codes	A
	U.S. Postal Service Postal Addressing Standards	Address format, state codes, street suffixes, secondary unit designators	B
Language	ISO 639, Codes for representation of language	Language codes	A
Race and Ethnicity	Office of Management and Budget Directive 15, Standards for the Classification of Federal Data on Race and Ethnicity	Race identification, Ethnicity identification	A
Occupation Classification	Bureau of Labor Statistics, Standard Occupational Classification System	Job activity classification	A
Vital Statistics	Centers for Disease Control (CDC) National Center for Health Statistics	Birth and death records, medical records, interview surveys, physical exams, laboratory testing, marriages and divorces, fetal death	A
	Bureau of the Census Current Population Survey.	Primary source of information on the labor force characteristics of the U.S. population. The sample is scientifically selected to represent the civilian noninstitutional population	C

[1] A - recommended for NCI use, B - recommended for further consideration by NCI, and C - standard is not recommended at this time.

Standard Type	**Standard Name**	**Standard Content**	**Recommendation**[1]
Measurement	HL7 codes for Units, Versions 2.X + (derived from the ISO 2955-83 standard (withdrawn by ISO in 2001) and ANSI X3.50)	Common units of measure, such as Celsius or mg/ml, intended to be combined with a numeric value to accurately express a result	A
	ISO 31, Quantities and units	Individual standards dealing with quantities in space and time, periodic phenomena, mechanics, heat, electricity and magnetism, electromagnetic radiation, chemistry, molecular physics, nuclear physics	A
Information Processing	FIPS 4-2, Representation of Calendar Date for Information Interchange	Means of representing calendar date to facilitate interchange of data among information systems	A
	ISO 8601, Numeric representation of dates and times	Formats for date and time	A
Education	United Nations Educational, Scientific, and Cultural Organization (UNESCO) International Standard Classification of Education (ISCED)	Designed to serve as an instrument suitable for assembling, compiling and presenting comparable indicators and statistics of education both within individual countries and internationally	C
	National Center for Education Statistics (NCES)	Works with federal, state, and local education agencies and researchers, to develop a series of data handbooks to provide guidance on consistency in data definitions for education data	C

Standard Type	Standard Name	Standard Content	Recommendation[1]
Health-related Vocabulary/Coding Standards			
	National Cancer Institute (NCI) Thesaurus	NCI reference terminology and description-logic ontology, providing comprehensive classification and characterization of types of cancer as well as cancer-related diseases, disorders, findings, abnormalities (cellular, molecular, and cytogenetic), gross anatomy, microanatomy, biological processes, genes, gene products, chemicals/drugs, combination therapies, mouse and other experimental models, and other topics	A
Basic Biology	Biological Pathways Exchange (BioPax)	Ontology for pathway information	B
	International Union of Biochemistry and Molecular Biology (IUBMB) and the International Union of Pure and Applied Chemistry (IUPAC)	Controlled vocabulary for nomenclature for biochemistry and molecular biology	A
Clinical	Common Terminology Criteria for Adverse Events v 3.0 (CTCAE)	Descriptive terminology for Adverse Event reporting	A
	Current Procedural Terminology (CPT) 4	Coding for evaluation and management, anesthesia, surgery, radiology, pathology and laboratory, medicine	B
	Healthcare Common Procedure Coding System (HCPCS)	Healthcare procedures, equipment, and supplies (Level 1) - used for Medicare billing Classification (national level) of physician and non - physician patient care services (Level 2)	B
	International Classification of Diseases for Oncology (ICD-O-3)	Coding for diagnoses of neoplasms - both topography and morphology - includes tumor location, cell type, tumor type, aggressiveness grade	A

Standard Type	Standard Name	Standard Content	Recommendation[1]
	International Classification of Diseases, Clinical Modification (ICD-9-CM)	Classifies diseases, conditions, symptoms, complaints/problems by diagnosis; supplementary classifications include health status, external causes of injury and poisoning, morphology of neoplasms, glossary of mental disorders, drug list numbers, industrial accidents and surgical, diagnostic, and therapeutic procedures	B
	International Statistical Classification of Diseases and Related Health Problems (ICD-10)	Collection, processing, classification, and presentation of mortality statistics	A
	Logical Observation Identifiers Names and Codes (LOINC)	Standard test names and codes, descriptive elements for other healthcare areas	A
	Medical Dictionary for Regulatory Activities (MedDRA)	Signs, symptoms, diseases, diagnoses, therapeutic indications, names and qualitative results, surgical and medical procedures, medical/social/family history, adverse event reporting	A
	Systematized Nomenclature of Human and Veterinary Medicine (SNOMED)	Findings/conclusions/assessments, procedures, body structures, function, organisms, substances, physical agents, occupations, social context/demographics, specimens, and other concepts	A
Genomics	Gene Ontology (GO)	Structured, controlled vocabularies describing gene products used for gene annotations	A
	HUGO Gene Nomenclature Committee (HGNC)	Controlled vocabulary of gene names and symbols for human genes	A
	Mammalian Phenotype Ontology (MP)	Standard vocabulary to describe phenotype data	B
	The Microarray Gene Expression Data (MGED) Society	Comprised of MIAME, MAGE, and the MAGE ontology, a suite of standards for microarray users and developers including an object model, document exchange format, toolkit, and ontology	A

Standard Type	Standard Name	Standard Content	Recommendation[1]
	Mouse Anatomy (adult – MA, and development – EMAP)	Ontologies used to annotate gene products	A
	Mouse Pathology (MPATH)	Ontology of terms used to annotate histopathology images	C
	Sequence Ontology	Ontology of terms for use in annotation of biological sequences	C
	Taxonomy	National Center for Biotechnology Information (NCBI) taxonomy of organism names represented in genetic databases	A
Drug Identification	National Drug File Reference Terminology (NDF-RT)	Drug classes, active ingredients (chemical structure), mechanics of action, physiologic effect, pharmacokinetics, therapeutic intent, commercial/clinical drug identification	A
	RxNorm Clinical Drug Vocabulary	Ingredients, drug components, drug formulations, drug strength representation, drug name synonyms, dosage forms	A
Health-related Transaction Standards and Models			
	Digital Imaging and Communications in Medicine (DICOM)	Standard method for the transmission of medical images and their associated information	A
Basic Biology	Systems Biology Markup Language (SBML)	XML exchange format for exchange of biochemical network models	A
	CellML	XML-based language for describing and exchanging models of cellular and subcellular processes	A
Clinical	American National Standards Institute (ANSI) X12	Insurance, claim payment, eligibility benefit inquiry, healthcare service coding, enrollment, encounter and claims	C
	American Society for Testing and Materials (ASTM) E1384-02a, Standard Guide for Content and Structure of Electronic Health Record	Patient identification (name, SSN, Birth Date, race, ethnicity), encounter, problem description, treatment, service type, testing (LOINC), appointment, provider information	C

Standard Type	Standard Name	Standard Content	Recommendation[1]
	Clinical Data Interchange Standards Consortium (CDISC)	Clinical trials data - general data (study name, protocol name, measurement units), study metadata (code lists), administrative data, reference data (lab normal ranges), clinical data	A
	Health Level Seven (HL7)	Patient tracking, scheduling, orders, results, clinical observations, billing, medical records, patient referral, patient care	A
	North American Association of Central Cancer Registries, Inc. (NAACCR, Inc.)	Demographic, tumor and staging, treatment and follow-up	A
Genomics	Biomolecular Sequence Analysis (BSA)	Specification defines a data model and interface format for data exchange on biological sequences	C
	Genomic Maps	Data model and structure for schema mapping for genomic maps	C
	Macromolecular Structure (Mms)	Specification for a data model and interface for exchange of macromolecular structure information	B
	PEDRo	Data model implemented in SQL and XML to support proteomics research	B
	Protein-Protein Interaction (PPI)	Data exchange format designed to bridge different formats of protein interaction databases	B
	Tissue Microarray (TMA)	Data exchange specification for tissue microarray data	A
Standards Review Bodies			
	Centers for Disease Control Public Health Information Network (PHIN)	Vocabulary and messaging standards; standards for data display and entry; standards for data transmission and management; implementation of applications and databases to support the adopted data standards	
	Consolidated Health Informatics Initiative (CHI)	Portfolio of existing clinical vocabularies and messaging standards enabling federal agencies to build interoperable federal health data systems	

Standard Type	Standard Name	Standard Content	Recommendation[1]
	Food and Drug Administration, Center for Drug Evaluation and Research (CDER)	Compilation of standardized nomenclature monographs that have been reviewed and approved by the CDER Nomenclature Standards Committee (NSC)	
	National Council on Vital Health Statistics (NCVHS)	Advises the government on recommended standards for adoption in the health care sector	

Exhibit 1. Standard Type Table

3.4.1 Common Demographic/Information Processing and Code Sets

As outlined earlier in this document, there are a number of organizations that have created demographic/information processing standards and code sets. The following section outlines recommendations for reviewing and adopting some of the standards previously discussed. Exhibit 1 identifies standards in each category. This section summarizes those standards that are most universally used.

3.4.1.1 Addressing Standards

The Universal Postal Union (UPU) has developed standard rules for the format and content of addresses for postal delivery. Rules and formats are provided for all countries that are part of the UPU. These rules are compatible with the mailing formats and coding systems supported by the U.S. Postal Service. It is recommended that these two standards be adopted for use by NCI for specification of mailing addresses.

The UPU standard does not allow for abbreviation of country names as part of mailing addresses. If there is a programmatic need for abbreviated country names, the ISO 3166 list provides a source on 2- and 3-character country names that should be used.

3.4.1.2 Language

ISO 639, Code for representation of language, provides an extensive list of languages, including the English name, French name, Local name, two character code, and name synonyms. It is recommended that this ISO standard be used for identification of languages.

3.4.1.3 Race and Ethnicity

The United States Office of Management and Budget Directive 15, Standards for Classification of Federal Data on Race and Ethnicity provides a list of race and ethnicity categories that are to be used by federal agencies and programs for reporting this type of information. It is recommended that NCI adopt this list.

3.4.1.4 Occupation Classification

The United States Bureau of Labor Statistics provides a list of Standard Occupational Classifications for job activities. It is recommended that NCI use these classifications for identifying occupations. The

Systematized Nomenclature of Human and Veterinary Medicine (SNOMED) also can serve as a source for occupation codes. Some coordination between the two standards may be needed.

3.4.1.5. Vital Statistics

The Center for Disease Control National Center for Health Statistics provides definitions for health related information terms and NCI should use these as a reference. These terms and definitions should be considered for inclusion in EVS. Terms are added to vocabularies under NCI control by the EVS project. EVS will request that terms be added to terminologies owned by other organizations upon NCI request.

3.4.1.6 Measurement

NCI should provide access to standard values for units of measure to facilitate information collection and analysis. ISO 31 provides a listing of quantities and units of measure. ISO 31 is available for purchase from the ISO or the American National Standards Institute (ANSI). NIST provides a source of measurement standards without charge (http://physics.nist.gov/cuu/index.html). CHI has recently adopted the HL7 codes for Units, Versions 2.X +, derived from the ISO 2955-83 standard (withdrawn by ISO in 2001) and ANSI X3.50. This standard can be used to define common units of measure, such as Celsius or mg/ml, that are intended to be combined with a numeric value to accurately express a result.

It is recommended that NCI determine the categories of measurement information required by program offices and make those standard units of measure available (for example, units of measure for volume, units of measure for mass, etc.)

3.4.1.7 Information Processing

FIPS Standard 4-2 and ISO 8601 provide for representation of calendar date and time. NCI should review these documents and adopt a standard way of representing calendar date and time for use in information exchange. HL7 should also be reviewed.

3.4.1.8 Education

NCI may have a need to designate educational levels of personnel. A couple of educational classification standards were reviewed but were not found suitable to meet NCI needs.

3.4.2 Health-related Vocabulary/Coding Standards

Standards for vocabularies need to support the required functionality, be easy to obtain and update, and to work well with the data provided by sources such as hospitals and other healthcare providers. The vocabulary standards have been subdivided by the following subject matter topics: Basic Biology, Clinical, Genomics. One clearly applicable standard is the NCI Thesaurus, a cross-domain translational terminology resource that covers all four categories of Health-related Vocabulary/Coding Standards. It is a CHI-recommended standard.

3.4.2.1 Basic Biology

Some standards for basic biology apply to NCI research. The initial recommendation is to use the International Union of Biochemistry and Molecular Biology (IUBMB) and International Union of Pure and Applied Chemistry (IUPAC) controlled vocabularies for nomenclature for biochemistry and molecular biology.

Other standards are emerging and NCI will need to stay informed to assess others of interest.

3.4.2.2 Clinical Code Sets

The NCI should look to EVS (including NCI Thesaurus and NCI Metathesaurus) as a source wherever possible. Note that NCI maintains vocabulary servers that robustly support the downloading of new vocabularies. Here are some initial recommendations for adoption of a consistent set of clinical vocabulary standards:

- Follow NCVHS recommendations by focusing on LOINC for laboratory test codes and SNOMED as the base vocabulary for clinical observations, e.g., results. Note, LOINC already provides codes for items that are asked on questionnaires. Explore using LOINC as a source for creating standard coding for items that appear on clinical trial forms.
- Evaluate use of SNOMED Clinical Terms (CT) for diagnosis codes, based on CHI recommendations. Some additional testing of these terms may be required to ensure they meet NCI needs. Where SNOMED codes do not exist to meet NCI needs, but are within the domain space of clinical medicine, new codes can be requested from SNOMED to facilitate mappings to hospital and provider clinical information systems.
- Use the Medical Dictionary for Regulatory Activities (MedDRA) and/or Common Terminology Criteria for Adverse Events (CTCAE) for adverse event reporting; use the International Classification of Diseases for Oncology (ICD-0-3) for severity classification of diseases.

3.4.2.3 Genomics

A number of standards are emerging in the field of genomics; recommendations for use are based on knowledge of current successful usage of the standards in the field. NCI will need to stay informed about other emerging standards. Initial recommendations for deployment in NCI programs include:

- Use the Microarray Gene Expression Data (MGED) suite of standards for gene expression data, including the object model, document exchange format and ontology.
- Use the Gene Ontology (GO) to describe gene products for gene annotations
- Use the HUGO Gene Nomenclature as a controlled vocabulary of gene names and symbols for human genes.
- Use the Mouse Anatomy ontologies to annotate mouse gene products.
- Use the National Center for Biotechnology Information (NCBI) taxonomy of organism names in genetic databases.

3.4.2.4 Drug Identification Standards

There are various drug coding "standards" within the pharmacy industry; yet none adequately address all the needs within the profession. Except for the National Drug Code (NDC) most of the "standards" are proprietary and are not utilized throughout the supply chain. The FDA issues the NDC standards to identify all drugs that are cleared for marketing in the United States. The NDC serves as a universal product identifier for human drugs. The current edition of the NDC Directory is limited to prescription drugs and a few selected Over-the-Counter (OTC) products. In addition, the NDC is not a comprehensive reference standard for medications to support a variety of clinical, administrative, and analytical purposes.

The emerging drug coding standards National Drug File Reference Terminology (NDF-RT) and RxNorm are developed to address the inadequacy of NDC and improve the interoperability of drug terminology; and are complementary to each other. Although both standards are still evolving and have not been widely used, it is recommended that NCI evaluate the adoption of both standards, as CHI has. CHI has adopted a set of federal terminologies related to medications, including the Food and Drug Administration’s names and codes for ingredients, manufactured dosage forms, drug products and medication packages, the National Library of Medicine’s RxNorm for describing clinical drugs, and the Veterans Administration’s NDF-RT for specific drug classifications.

NCI EVS works with the VA through an interagency MOU on the cooperative development of NDF-RT. Information on cancer agents is being developed in the NCI Thesaurus using a common model to be shared with VA. Part of the VA data has already been incorporated into NCI Thesaurus and Metathesaurus. Furthermore, it is critical for NCI to remain involved in the development process of both standards to address the needs of clinical trials.

3.4.3 Health-related Transaction Standards

One recommended biomedical transaction standard that does not fit into one of the categories below is the Digital Imaging and Communications in Medicine (DICOM) standard, jointly developed by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) as a standard method for the transmission of medical images and their associated information. DICOM is used or will soon be used by virtually every medical profession that utilizes images within the healthcare industry and is applicable to the exchange of cancer imaging information.

3.4.3.1 Basic Biology Transaction Standards

The Systems Biology Markup Language (SBML) and the related Cell Markup Language (CellML) were considered to be useful exchange formats for exchange of molecular physiology information.

3.4.3.2 Clinical Transaction Standards

NCI supports a wide range of clinical trials and receives data about the progress of trials and about their outcomes. The reporting of this data is not carried out through standards-based messaging, however such a report has a lot in common with a message. The report has a defined hierarchical structure, and its contents have to be stored and, ideally, compared to the reports of related studies. The NCI should set the goal of standardizing this reporting, and should select the messaging standard or standards that support this goal.

This report has described two messaging standards that could be used, Accredited Standards Committee (ASC) X12, and HL7. Of these, HL7 has the deepest base within the clinical arena, and includes a direct focus on clinical trials through the Regulated Clinical Research Information Management (RCRIM) Technical Committee. HL7 has also put a good deal of effort into ensuring that its messaging works well with clinical vocabularies—which is a key concern for NCI. As a result, it is recommended that NCI focus and work on HL7 as its messaging standard.

It is important to note that HL7 publishes two sets of messaging standards, Version 2 and Version 3, which display significant and relevant differences. HL7 Version 2 contains the original set of HL7 messages that have been worked on, and widely implemented over the last 15 years. In particular, HL7 Version 2 is the standard that U.S. hospitals and laboratories use for recording patient encounters, clinical orders, and test results. While Version 2 supports adverse event reporting[2], it does not include messages devoted specifically to clinical trials. On the other hand, HL7 Version 3 is based on a reference model that could be mapped to repository databases at NCI, and many of the Version 3 messages that will be needed by NCI are in the process of being defined within the Version 3 ballot packages. Additional

[2] This part of the Version 2 specification has not been broadly implemented.

specifications could be created by NCI using the published messaging methodology. Furthermore, projects working within the context of the CDC-sponsored Public Health Information Network (PHIN) have chosen to implement Version 3 messaging in various contexts for which there is a need for public health-related specifications.

Also, it is worth noting that the HL7 Reference Information Model (RIM) can be used as a point of departure for static data, for classes, and data elements to be a basis for database design. While the model is highly generic, it can be specialized as needed. Using the RIM makes it easy to tie in the data structures created with the messages and vocabularies sponsored by HL7.

Given this situation, where the developing standard has important features that are relevant to NCI, but the existing version is in common use and will continue to be used for the near future; it is important for NCI to maintain the ability to work with both HL7 Version 2 and HL7 Version 3. The following recommendations apply to these two families of HL7 messaging.

Version 2 Strategy: The widespread use of Version 2 should be recognized by supporting Version 2 messaging in situations in which a) data is received from healthcare providers, and b) Version 2 specifications are available and widely used. This implies the continued use of Version 2 messaging for clinical (most notably laboratory) results and for the messaging fundamental to healthcare data processing, e.g., order entry, Admissions, Discharge, Transfer (ADT) management. NCI should take the perspective that these transactions will not be migrated to Version 3 until it is widely adopted by U.S. healthcare providers. That will not happen for a number of years.

Version 3 Strategy: Version 3 provides users with specifications built by reference to an explicit data model, whose structure and semantics are both clearly exposed and easy to map to a RIM-compliant database. For these and other reasons, Version 3 should be the version of choice when developing new message specifications, and for transactions that flow between NCI and governmental departments and agencies.

Common Elements and Vocabulary: It is important to note that the key issues of vocabulary, object identification, and message parsing have to be addressed in both types of HL7 messaging. It is recommended that NCI use common vocabularies and schemes for object identification in both Version 2 and Version 3. Consideration should also be given to implementing XML encoding for the Version 2 messages in place of the original HL7 proprietary delimiter based encoding.

3.4.3.3 Genomics Transaction Standards and Models

A number of specifications are emerging that define models or data exchange specifications for genomics and proteomics data. Only a few were considered to be sufficiently mature to recommend for near-term deployment.

Development of the Tissue Microarray (TMA) data exchange specification for tissue microarray data was sponsored by the NCI, and it is planned for use within NCI programs.

Proteomics standards were generally considered insufficiently mature to adopt at this time.

4.0 DATA STANDARDS APPENDICES

In the appendices that follow, each data standard is described by: sponsoring organization, current and historical versions, usage, applicability to NCI, potential for management in the caDSR or the EVS (and related maintenance issues), and potential for NCI involvement in standard development. Standards will include international, national, federal, and those sponsored by other organizations. Within each appendix, standards are categorized into the three groups defined in section 1.3 - demographic/information processing, vocabulary, and transaction standards. Vocabulary and transaction standards are further subdivided by broad subject matter categories - basic biology, clinical, and genomics. Appendix A contains the standards considered most relevant to NCI and suitable for adoption. Appendix B contains those standards potentially suitable to NCI and possibly worthy of further consideration. Appendix C contains descriptions of some standards that were reviewed but found to be redundant with other, more widely used standards, not applicable to NCI, or insufficiently mature to consider at this time. Appendix D describes the relevant standards review and approval bodies that do not develop standards but review them for adoption or approval.