Data Standards in Clinical Research: Gaps, Overlaps, Challenges and Future Directions

RACHEL L. RICHESSON, PhD, MPH, JEFFREY KRISCHER, PhD

Abstract  Current efforts to define and implement health data standards are driven by issues related to the quality, cost and continuity of care, patient safety concerns, and desires to speed clinical research findings to the bedside. The President’s goal for national adoption of electronic medical records in the next decade, coupled with the current emphasis on translational research, underscore the urgent need for data standards in clinical research. This paper reviews the motivations and requirements for standardized clinical research data, and the current state of standards development and adoption—including gaps and overlaps—in relevant areas. Unresolved issues and informatics challenges related to the adoption of clinical research data and terminology standards are mentioned, as are the collaborations and activities the authors perceive as most likely to address them.

Introduction
Efforts to build a national health information infrastructure (NHII) and supporting data standards must address the needs of clinical research.1 Clinical research, as defined by the National Institutes of Health (NIH) is patient-oriented research conducted with human subjects (or on material of human origin that can be linked to an individual).2 Clinical research includes investigation of the mechanisms of human disease, therapeutic interventions, clinical trials, development of new technologies, epidemiology, behavioral studies, and outcomes and health services research. The broad scope of clinical research, coupled with the infusion of technology, has generated increasing amounts of data, and the scientific community needs to identify strategies to share it in meaningful ways. The NIH policy on the sharing of research data3 is bringing forth questions about how data should be represented for data sharing, and making the need for clinical research data standards critical and immediate.

Data standards are defined here as consensual specifications for the representation of data from different sources or settings. Standards are necessary for the sharing, portability, and reusability of data.4–7 The notion of standardized data includes specifications for both data fields (variables) and value sets (codes) that encode the data within these fields. Although the current data standards focus is on regulated research (often the narrower context of clinical trials) and their business activities (e.g., safety reporting, study reporting to regulatory bodies), it is important to mention that clinical research includes many other types of research, including observational, epidemiological, and outcomes research, as well as molecular and biology research (e.g., genetics and biomarkers for disease). Although important, this discussion does not address the “-omics” standards,8 but rather clinical, laboratory, procedure and observation data collected in the context of clinical research subject visits.

The permeation of clinical research data standards that are harmonious with clinical care standards is required for the sharing of patient data between healthcare and research—one ambition of the NHII.9 The goals for the NHII include the seamless integration of clinical research data to/from patient care data to/from population data and existing medical knowledge bases, making standardized data in clinical research a high priority.5,9,10 Interoperability between healthcare and clinical research data can create opportunities for increased subject enrollment, evidence-based medicine, and population monitoring. This paper describes data standards requirements for subject data in the clinical research domain, the nature of overlaps and gaps in current standards coverage, and highlights key informatics challenges that remain.
Current Activities Related to Clinical Research Data Standards

There are several stakeholders (Appendix 1 available as an online JAMIA supplement at www.jamia.org) for the uniform adoption of data standards in clinical research including both pharmaceutical companies and regulatory agencies such as the U.S. Food and Drug Administration (FDA) and its international equivalent the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). Other stakeholders include academic and government-based clinical researchers, and government agencies that are large sponsors of clinical research in specified domains (e.g., National Cancer Institute, National Heart, Lung, and Blood Institute) and those NIH institutes that sponsor clinical research agendas (e.g., National Center for Research Resources, NCRR). Because of the importance of data standards in healthcare delivery and federally funded research, various NIH institutes have standards efforts and dedicated resources for examining data standards within their domains of interest. The NCRR is exploring the impact of current government standards initiatives upon broad clinical research interests expressly.11

In 2003, all federal agencies with health data interests committed to adopt a set of voluntary data standards recommended by the Consolidated Health Informatics (CHI) initiative, a U.S. based multi-agency movement for healthcare data standards.12 Their recommended standards for over 20 areas of healthcare data have been endorsed by the Department of Health and Human Services’ Office of the National Coordinator for Health Information Technology (ONC).13 While the CHI standards defined in 2004 have yet to be widely implemented in the commercial world, government agencies, such as Centers for Medicare and Medicaid Services (CMS) and the Veteran’s Administration (VA), are striving to incorporate them. Some CHI standards, such as HL7, were well-used prior to being named CHI standards, while others, such as SNOMED CT, were not widely implemented and their adoption in federal healthcare activities is moving more slowly. The CHI standards identification teams of 2004 postponed recommendations on many key data areas, including physical exam, medical history, and adverse events, and it is not clear if the ONC’s Health Information Technology Standards Panel will continue with that agenda. Clinical research interests were not specifically represented in the CHI activities of 2003-4, but advocates for clinical research needs in health IT data standards do have a formal presence in the current ONC activities.

The National Library of Medicine (NLM) operates in a leadership and coordination role in the identification of an interlocking set of standards that would completely address all NHII data representation needs. The NLM has developed and maintains knowledge sources and tools to facilitate access to standard terminologies as well and to coordinate their use.14 In 2003, the NLM procured a public U.S. license for the use of SNOMED CT via their Unified Medical Language System (UMLS). The UMLS is a multi-purpose resource that includes concepts and terms from over 100 different source vocabularies, and establishes linkages across these source vocabularies with its semantic network representation of important concepts and relationships in the biomedical domain.15,16 The UMLS tools support the mapping of multiple data standards to a common set of concepts via the Metathesaurus,17 and the NLM historically functions in a coordinator and funder role in creating needed mappings across heterogeneous data standards. Additionally, the NLM has been supporting the development of tools for the access and use of CHI standards, and for achieving interoperability between them. They have funded LOINC and special projects that facilitate government agency interests in standards, including HL7. Additionally, NLM has developed and delivers RxNorm, a database of drug concepts, including the clinical drug (drug + route + dose) which is being linked with the FDA’s new structured product labels (i.e., package inserts) effort.

The ICH is a collaboration of the regulatory authorities of Europe, Japan and the United States that formed in 1990 with the aim of harmonizing scientific and technical aspects of product registration in order to eliminate the need for duplicate testing in the development of new medicines. The ICH has developed multiple reference models for the clinical research domain and the E2B specification of data elements for transmission of Individual Case Safety Reports of all types of individual case safety reports, regardless of source and destination.18 The E2B data model standard is the foundation that other clinical research data standards groups are drawing from to support development of standardized electronic regulatory data reporting applications in the short term.

The dominant discussion forums for moving toward clinical research data standards that support applied uses are the Clinical Data Standards Interchange Consortium (CDISC) and the Regulated Clinical Research (RCRIM) Technical Committee of Health Level Seven (HL7). These groups are very different in terms of membership, organization, and purpose. The CDISC membership is dominantly from large pharmaceutical companies world-wide, but also includes the FDA as well as representation from governmental agencies such as the VA and NCI. The immediate goal of CDISC is to create standard data models for regulatory submissions. While CDISC has commendable industry participation and motivation, it is not a formal standards development organization, and there is a risk that the organization might not address the needs of all stakeholders.

HL7 is a not-for-profit volunteer organization dedicated to produce standards for clinical and administrative data in all health settings, and is an American National Standards Institute (ANSI)—accredited Standards Developing Organization (SDO).19 Like all ANSI-accredited SDOs, HL7 adheres to a strict and well-defined set of operating procedures that ensures consensus, openness and balance of interest. The Regulated Clinical Research Information Management (RCRIM) Technical Committee shares some of the same goals as CDISC, but also represents broader clinical research and patient safety interests. Because HL7 also addresses more stakeholders than just clinical researchers, the time for discussion and approval of the standards can be lengthy. This formal and required discussion and approval of developing standards throughout HL7 increases the likelihood that the standards that are defined in technical committees and special interest groups such as the RCRIM, will be
Both CDISC and HL7 are developing models for messaging or transferring data (e.g., records, reports, or data sets) to drug research regulatory organizations. The CDISC effort focuses on building formal data models for regulatory reporting (to FDA). Although current (2.x) versions of HL7 use relatively simple models to support messaging in clinical settings, the HL7 version 3 (not widely implemented) relies upon a very abstract information model, the Reference Information Model (RIM) that is broad and flexible enough to address any messaging need in the healthcare domain. CDISC is developing more practical data models (designed for very narrow, explicit regulatory needs), with a finite set of variables needing controlled vocabulary. Both HL7 and CDISC have terminology teams or liaisons with terminology groups tasked to put the terminology (or controlled value sets) in the slots of the information models that they have developed, but the extensive differences in the HL7 and CDISC models raise the concern that “ideal” terminology standards for each application might differ between the groups. Heterogeneity across HL7 and CDISC models also invites the risk that the same terminology standards can be applied differently in different applications. In addition to terminology issues, the harmonization of HL7 and CDISC standards for message/information structure, encoding of data types, and communication conventions are imperative. CDISC and HL7’s RCRIM groups began formally working together in 2001 and are committed to achieving syntactic and semantic interoperability between their standards. It is not clear whether the CDISC data models and the broader HL7 messages will ever be synchronized, but there is promising work in harmonizing CDISC lab reporting standards with HL7 (version 2 and 3) messages, including the use of LOINC test codes within both models.

The incentive to harmonize the CDISC and HL7 models is strong, although the task is daunting due to the differences in complexity, conceptualization, and in the levels of abstraction between the models. The Biomedical Research Integrated Domain Group (BRIDG) formed in 2005 solely to link the CDISC data reporting models with the HL7 RIM. The (BRIDG) model is a domain analysis model of protocol-driven biomedical and clinical research, developed to provide a comprehensive conceptual model of the clinical research domain as a basis for harmonization across information model standards. The domain model is the result of work from HL7 RCRIM, CDISC, NCI, and the FDA. The BRIDG Model has recently been formally adopted by both CDISC and the HL7 (RCRIM) Technical Committee as their domain analysis model and it is supporting National Cancer Institute’s (NCI) cancer Bioinformatics Grid (caBIG). The BRIDG model is intended to be the conceptual backbone to which all CDISC and HL7 RCRIM implementation models link, thereby creating interoperable applications within and across both organizations. Proof of concept and pilot demonstrations of this harmonization are in early development.

Although not a formal SDO, the NCI has been building a standards infrastructure for years, offers robust terminology resources, and is an important and active participant in most clinical research data standards venues. The NCI has developed the Common Terminology Criteria for Adverse Events (CTCAE), a standard for adverse events (AE), which is perhaps the most comprehensive AE classification for general clinical research, despite its origins in oncology. The NCI has created a strong infrastructure for terminology maintenance, mapping, and access activities. The NCI serves as a host for CDISC data elements, value sets, and terminology as they are being defined. In addition, the NCI is hosting controlled terminology for the FDA. The NCI has been successful at standardizing (to some extent) clinical research data within its many sponsored studies, although they too are struggling with the standards gaps and overlaps we describe in the next section.

**Status of Data Standards in Clinical Research**

One of the greatest challenges in identifying data standards for the clinical research domain is to reconcile the requirements of varied investigators and data users with the need for common standards. While individual investigators might only want a subset of an information model or a standardized terminology, collectively clinical research needs terminologies with comprehensive coverage, including broad concepts (e.g., “abnormal nervous system finding”) and detailed ones (e.g., “abnormal lateral conjugate gaze”). Further, clinical research workflows create requirements for representing important nuances of clinical and research data, particularly in regulatory contexts, which differ from other healthcare areas. For example, although AEs are clinical findings and observations, they are defined in the context of a research protocol, imply subject participation in a study procedure (usually intervention), and contain other semantic dimensions (e.g., attribution, expectedness).

**Clinical Research Data**

The clinical research domain, as a whole, includes data from the spectrum of broad constructs shown in online Table 1. If there is a current U.S. standard, the standard and the organization naming the standard are listed. We characterized whether each construct had a gap or an overlap in named U.S. standards. Competing standards are listed for the constructs with overlaps, and potential standards (with at least some relevant content) are listed for the constructs with gaps in standards coverage. Because standardized data includes specifications for both data fields (variables) and value sets (coding systems), the state of standards adoption is described by both variables and value sets for most constructs examined.

**Constructs**

Conceptualizations of the clinical research domain (i.e., the definition and organization of constructs) can be based upon data content types (e.g., clinical observations, reported symptoms, diagnoses) or by workflow activities or artifacts (e.g., physical exam, medical history, adverse events). In Table 1, we present constructs that are related to both data content types and workflow activities in order to mesh with how other groups such as CHI conceptualized health care data domains for endorsing standards. We present Table 1 as a starting point to visualize broad target areas for data standards in the clinical research domain, and to illustrate the number of data areas that are missing named standards (“gaps”) or have one or multiple competing standards (“overlaps”).
Gaps

While some clinical research constructs are covered by existing data standards, there are important areas that have gaps in standards coverage. These include study descriptive information (including study design and randomization/blinding features), subject and provider identifiers, study eligibility criteria, subject disposition descriptors (e.g., eligible, enrolled, loss to follow-up), study events (e.g., baseline visit, follow-up visit, physical exam), protocol deviations, medical device names, and vital signs. Although data standards for medical device names and vital signs are relevant to broader healthcare interests, the majority of gaps in clinical research data standards are those that are very unique to clinical research. This is understandable considering that the CHI initiative primarily was focused on naming data standards for health care delivery.

For constructs with standards gaps, some (subject identifiers, protocol deviations) have just one or a few candidate standards and no apparent de facto standards with a broad user base in clinical research. Other constructs with formal standards gaps (study descriptors, demographics, subject disposition, medical devices, and vital signs) have multiple candidate standards and overlap of standards is likely in the future. Another gap, listed as a separate construct in Table 1 but relevant to all construct areas, is needed standards for encoding missing data—e.g., unknown, not reported, not assessed, refused, etc. While this has been addressed by the “flavors of null” work within HL7 version 3, it needs to be simplified and adopted by data collection and management in the clinical research domain.

There is a conspicuous lack of named standards for the structuring of questions and case report forms, particularly in the areas of physical exam, medical history, family history, and eligibility criteria. It is important to note that while named standards (i.e., SNOMED CT and MedDRA) exist for the content of these activity areas, standards for how they are used (metadata and question modeling) are both important and lacking. Such standards would support a consistent structure and use of standardized terminologies in a variety of applications, including interfaces to electronic health records, public health questionnaires, message models such as HL7, or clinical research case report forms. The importance of question-level metadata is evidenced by applications such as the NCI’s Cancer Data Standards Repository and CDISC’s Clinical Data Acquisition Standards Harmonization (CDASH) project—a new project focused on the development of data standards in case report form design. The importance of the exchange and reuse of federally-required patient/client assessment and other functioning and disability content to Centers for Medicare and Medicaid Services (CMS), and other U.S. Department of Health and Human Services (DHHS) agencies, resulted in placing this area on the CHI agenda. The CHI standard (Fall 2006) for the area of Functioning and Disability includes a combination of data standards, including LOINC, to represent the items and batteries of items on standardized federally-required patient/clínical assessments and other disability content across the federal healthcare enterprise. The recommendations were preceded by a DHHS-sponsored work group that looked at many of the questions and determined that the majority of clinical (health related) content (which the team called “usefully related” content) was covered to some degree by relevant vocabularies such as International Classification for Functioning, Disability, and Health (ICH) and SNOMED CT, but that structural features of questions, including answer groups, were not represented well by recommended vocabularies.

The feasibility of LOINC to represent items in standardized questionnaires has been demonstrated. While the LOINC model is useful for capturing key features of a question, the usefulness of LOINC for indexing questions from standardized instruments would improve with the inclusion of hierarchical knowledge, and the inclusion of additional relevant question attributes, such as the exact item wording, that can influence its appropriate use and analysis. Lacking in the LOINC model is necessary comprehensive hierarchical knowledge; thus the recent CHI standard for Functioning and Disability data recommending the use of LOINC plus controlled terminology such as SNOMED CT in the context of representing assessment questions.

Overlaps

The constructs labeled as overlaps on Table 1 are those for which there are multiple standards named by one or more authoritative bodies. Several candidate data standards (which can be considered potential de facto standards) that have content to address these constructs are listed in the final column. Even in areas with a named U.S. standard, we do present potential competing standards in the last column with the intention to illustrate, collectively, the scope of the coverage overlap across a broad range of clinical research data constructs. We also present the domain/scope/intended use of the potential existing standards to show that, though potential standards exist, effort may be required to make them useful for application in the clinical research domain. Lastly, we provide a column with our subjective and approximate estimate of the amount of work required to achieve standardization—minimal, moderate, or significant. Focused research of potential existing standards in each construct area will be necessary to complete this list. This early assessment of existing data standards is a reasonable first step for standards identification efforts in areas where there are no named standards. There is no way to know how many additional locally developed value sets or terminologies exist for any of the constructs on Table 1 but the extent of the potential overlap is certainly much greater than that depicted here.

The overlap of multiple named standards is most prominent in the (value set) area of Physical Exam observations and findings. The U.S. CHI initiative has identified SNOMED CT as the standard for several areas relevant to patient physical exam data in clinical research—problem lists, diseases, and anatomy, and the FDA recently named SNOMED CT as a standard for prescribing information in the Structured Product Labeling effort. (The FDA posts a subset of SNOMED CT codes to be used as the problem/finding in their new labels.) However, the ICH has endorsed MedDRA as a standard for all clinical data since 1991, and MedDRA is embedded into the workflow and information systems for a majority of pharmaceutical companies. Despite U.S. public access to SNOMED CT via the NLM since 2003 and the adoption of SNOMED CT by 9 countries, licensing issues remain a
barrier for global use, and the effect of the International SDO status of SNOMED CT has yet to be seen.

Hidden overlap might still exist in areas where standards are defined, but lack of guidance or implementation experience makes the boundaries between related standards ambiguous. For example, the CHI recommended standard for Allergy Data consists of a suite of standards for different types of allergies (e.g., drugs and biologics, food substances, device-related substances, environmental toxins). Practical implementation of these various standards might reveal overlap at the boundaries between the standards, as the definitions for food and medications are often not clear.\(^{26}\)

To give the reader an appreciation of broad clinical research areas with named data standards, the constructs presented in Table 1 are deliberately collective in nature. Some areas embody multiple elements, each of which could be explored deeper for standards coverage, where additional gaps and overlaps would very likely become evident. For example, the Demographics area consists of variables such as race, ethnicity, gender, education, occupation, and income. The CHI has recommended OMB standards (the White House E-Gov initiative) for race, ethnicity, occupation, and industry, but other demographic constructs, such as education and income have no named standards (i.e., these areas represent standards gaps) but have multiple competing de facto standards within the U.S. government and external research areas, revealing potential for future overlap.

The selection of data domains presented in Table 1 was heavily influenced by CHI efforts at defining standards areas. However, the CHI effort’s conceptualization of areas was broadly focused and does not easily translate into defined healthcare or clinical research processes and data flows. The early definition of CHI domains was largely driven by taking inventory of areas where current data standards exist—although those standards broadly cover many types of clinical research data and artifacts (e.g., SNOMED CT covers data areas such as anatomy, findings, etc. that are in medical history and physical exam reports, and HL7 messaging standards could apply to laboratory reports, safety reports, etc.) In addition to our desire to include all applicable CHI standards in our presentation, we also attempted to include constructs for key clinical research activities. This was somewhat data-driven based upon our experiences of types of data collected in the variety of clinical studies that we support. Clearly, a terminological approach to standards inventories (such as the CHI approach) carries a danger of not addressing different coding requirements which might apply to the same data construct in different business processes (e.g., the specification for ordering laboratory tests might not be the same as that required for receiving results), and future elaboration and expansion of this table is justified. The lack of a unified conceptual model for understanding healthcare and research processes and data constructs complicates any needs assessment for standards. Domain analysis models (e.g., BRIDG), once they are complete, can and should inform the constructs in Table 1. It is likely that additional gaps and overlaps are present but not yet realized.

**Future Directions and Informatics Challenges**

There are both research-specific and informatics-specific issues involved in achieving data standards, and individuals with training in both disciplines will be indispensable. A principal challenge is the lack of explicit and consensual understanding of the constructs, activities and inter-relationship flows of clinical research data. Once there is consensus on exactly where standards are needed, one can more precisely identify persistent gaps and overlaps, and critically evaluate potential terminological standards. We outline specific challenges for achieving data standards in clinical research practice, and suggest strategy and leadership, in order to guide future discussion within the clinical research and informatics communities.

**Lack of Definition of Purpose for Data Standards in the Clinical Research Domain**

An explicit and consensual understanding of the intended nature of data sharing will further illuminate the gaps and overlaps of current named standards, and dictate in which standards activities clinical researchers need to be represented. Lobbying efforts to bring forward clinical research data needs to relevant standards bodies are warranted and should continue, but a sense of purpose on behalf of the clinical research community can direct these discussions toward practical and worthwhile applications and demonstrations. Arguably among the most successful efforts at standardization efforts are the NCI and CDISC, which work to improve efficiency of specific business or regulatory processes. To drive the identification of appropriate data standards for broader clinical research needs, the intended uses for standardized data must be defined. Potential drivers, including the NIH data sharing policy\(^3\) and use cases for interoperability of health delivery and clinical research data, should be explored and exploited for hastening standards progress.

There is palpable tension in the clinical research data standards community between achieving tangible solutions for real business problems, and long-term interoperability that bridges among the broader research community, the broader healthcare community, and the global community. This tension between short-term progress and long-term vision has long been, and continues to be, an issue for the adoption of data standards in the context of electronic medical records and national healthcare information infrastructure. Rather than be discouraged, the clinical research and informatics communities should seek to understand and learn from the challenges, failures, and successes from over forty years of experience between information technology and healthcare delivery.\(^{27}\)

We recommend that the AMIA membership, particularly the Clinical Research Informatics (CRI) Working Group, develop use-cases for the sharing of data across clinical care and research applications. These use-cases should identify situations where data sharing might occur. Additionally, the CRI Working group should clearly delineate which types of clinical data, and under which circumstances, might have the rigor and precision and reliability to be used for research purposes. Clinical research data standards dialogs ultimately require representation cognizant of the specific data standards needs.
all types of clinical research. We recommend that the NIH take a major role in defining purposes for the sharing that will then drive the standards requirements. The many institutes and components of the NIH represent a breadth of research foci and goals, and collectively are a major funder of clinical research activities worldwide. Key stakeholders within NIH would include those with broad research interests, such as NCRR, and those from disease-specific components that have extensive research agendas. The collaboration of the NLM, which is familiar with data standards development and adoption issues (including sophisticated terminology interactions) that have surfaced from the clinical care arena, will be an asset. The CTSA activities might prove to be a coordinator for dialog and coordinated representation of clinical research interests—essentially functioning as a nationally-driven clinical research data standards task force.

Information Model Selection and Terminology

Implications

Variation in data models across industry and emerging “standard” information models complicate efforts to identify terminologies that are ideal in multiple information models, and imply the need for specifications for which parts and how terminological standards should be used. Terminologies with complex terminology models, such as SNOMED CT, can have multiple options for concept representation and have resulted in the need for guidance on how the terminology should fit into an information model. Do you insert the terminology concept(s) “left arm” or “arm” + “left” in an information model with data fields for both “body site” and a “laterality?” Does this compare with data encoded in a different information model with a single data field called “body site?”] Additionally, terminologies whose scope is bigger than that of the information model can create the need for boundaries as to which parts of the terminology will be used in a given information model.28

The issues with information model—terminology interactions have been discussed for some time5,29–32 and are central to achieving practical data standardization. The problem is most notable with HL7 RIM and SNOMED CT—both with very sophisticated and comprehensive models. This issue has been the subject of several years of focused activity on the part of the TermInfo working group of the HL7 Vocabulary Technical Committee.35,36 Typical terminology evaluation studies take place in controlled contexts, though the authors are not aware of any analyses of coding consistency that control for the dynamics of the terminology and information model interaction. There is no single unified information model to support clinical research needs. If multiple information models are inevitable, then strategies (e.g., BRIDG) for the harmonization and co-evolution of these models will be necessary and should be pursued.

We believe that, despite the years of work that both HL7 and CDISC have invested in developing their standards, the models are still new and opportunities for their harmonization are real and available. We support the continued development of the BRIDG domain model by the current stakeholders (HL7, CDISC, FDA, and NCI) and encourage its evaluation as a means to harmonize emerging application models from all of those organizations. In addition, we believe that other NIH institutes and components, particularly those with broad spectrum of research interests, such as the NCRR, should participate and share broader clinical research perspectives. With representation from both public and private research interests, broad domain models such as BRIDG might be a means by which heterogeneous models might co-evolve and become complementary.

Because of the SDO status and broad scope of HL7 mission and membership, we propose that balloting and maintenance of CDISC standards be formally managed through HL7, with CDISC operating as a consensus group for the regulated research community.

It is likely that the terminology/information model interactions have been underestimated by all stakeholders, and are a potential danger to achieving standards in the clinical research domain, despite the commitment to harmonization efforts from clinical research stakeholders. Terminology should be considered at the stage of model development, and revisited often. We propose increased collaboration between terminologists and domain experts from all stakeholder organizations regarding the semantic coordination of models and terminology for all projects. In key areas, such as clinical findings and adverse events, the issues of competing terminologies must be addressed and resolved concurrent with the development of information models. Because potential terminology model—information model interactions are so important, we consider that the active involvement of NLM in the modeling and development of information model standards will be invaluable, as their expertise in terminologies could help predict and attenuate variation in terminology implementations across competing standards.

Lack of Quantitative Evaluation of Competing Terminologies

Gaps in data standards will have to be filled by either extending existing standards or building new ones. Where there are currently overlaps in coverage, it will be important to have operational criteria that facilitate objective comparisons of competing data standards. To make informed decisions about best practices, decision-makers need comparative data, including evaluative studies needed on which is “best” in a given domain for a given purpose. In all likelihood, more than one of the candidate standards would be satisfactory, so ranking of evaluation criteria should allow for objective comparison of competing data standards. There have been few studies that actually examine the nature, scope and depth of clinical research data. To date, coverage is a critical evaluation feature, but other issues, such as organizational and usability issues must be considered.35,36 The ranking of evaluation criteria would vary by task, but broad clinical research data standards should weigh international suitability, access, and maintenance of terminologies as high as content coverage and other desiderata. We encourage the terminology research community and clinical research community to identify and expand quantitative measures for evaluation (including new requirements unique to clinical research data).

Despite the lack of acceptance of a single standardized information model, or of multiple harmonized models, within the clinical research community, various organiza-
tions, including CHI, have named terminological standards for clinical care data. In general, the evaluation of these standards for clinical research is less than straightforward, because they encompass a broad range of constructs, are designed for different purposes, are expanding to address more needs, and have heterogeneous structures and various levels of granularity. An intuitive strategy for achieving data standards in clinical research is to decide on the information model first, and then select terminology or terminology sub-sets that are appropriate for data instance representation within the model. This top-down approach has been distracted by concurrent CHI initiative to name terminological data standards for certain knowledge areas (e.g., problem lists, anatomy) whose fit into the real world applications and data models is unclear at this time. The existence of terminological standards in the absence of information model standards has created confusion for implementers as the application of terminological standards is dependent upon the information model. An additional risk is the non-standard use of standard terminologies – especially as multiple implementation models are introduced within and across organizations such as CDISC and HL7.

A related issue is developing on-going models for collaboration and maintenance of data standards, so that today’s harmonization of competing information models and associated terminology is not lost tomorrow. Regular and coordinated communication between standards groups can facilitate the co-evolution of models and data representation for clinical research data that can reduce or even eliminate heterogeneity going forward. The use of terminology sub-sets and transformations (e.g., maintaining terminology sub-sets outside of the terminology developer, adding abbreviations, definitions, etc.) must be carefully monitored by knowledgeable stakeholders from HL7, CDISC, NCI, and NIH, so that use of terminological standards and value sets occurs uniformly across organizations. We propose the formal communication between policy makers and terminology experts from HL7, CDISC, NCI, and NIH to agree on high level processes for communication between CDISC and HL7 standards activities and developments. We think that this communication should include information-model – terminology interaction and should focus on seizing opportunities for harmonization and co-evolution of the standards. Although not a direct funder or stakeholder for clinical research, the NLM would be a vital party to objectively identify situations where CDISC and HL7 are using terminological standards in ways that might impede future interoperability.

Technology Needs

Technological needs for achieving data standards include solutions for human users to access and view the vast content and heterogeneous structures of complex information models and terminologies. Tools and resources that illustrate competing information models, in relation to concrete tasks and well-defined work processes, are needed. Tools are needed so that evaluators easily can visualize terminology structures, easily search for needed concepts, and easily realize any interactions between the terminology standard and the information model supporting their applications. Tools to bridge the divide between terminology and context-dependent sub-sets, including mappings between terminologies—also subject to change and updates—will be relevant to specific research needs.

The use of data standards at the point of data collection necessitates technologies to facilitate the storage and retrieval of clinical research questions and answers, and to relate them to controlled terminologies. Applications such as the NCI’s caDSR, built from ISO specifications, that relate terminological concepts to question-answer sets common in clinical research data collection, are promising demonstrations of tools needed for the reduction of data variation and the permeation of data standards within clinical research organizations.

The continuous infusion of technology into the clinical research workspace, as well as high level efforts at re-engineering and streamlining current clinical research practice, are changing research activities and workflow. Tools that support communication and collaboration across clinical research interests can enhance the community’s ability for proactive discussion about dynamics in both clinical research practice environments and data standards worlds. Certainly, the desire for tools that aid the analysis and exploration of shared data will continue to grow, and their development and use might bring this effort full circle and demonstrate value of data standards and shared knowledge, within and across various domains and settings. It is our hope that these tools will evolve naturally from a variety of stakeholders as the importance of the outstanding issues we raise here become clear.

Other Challenges

The problems of integrating U.S. data standards in international settings foreshadow more potential standards overlap in the future. Health Level Seven (HL7) has seen the issues related to inappropriateness of some U.S. data standards (particularly race) for international uses and has created “realm-specific” code sets that essentially allow different value sets for different countries. The comparability and interoperability of these distinctions remains to be seen. The international scope of big pharmaceutical companies, coupled with enabling technology for multinational research participation, accentuate the relevance of global perspectives for clinical research data standards.

The importance of terminology-related metadata that can assimilate heterogeneous coding systems is becoming an important research and development area of relevance to addressing construct areas with overlaps. The UMLS and the NCI Thesaurus embody an underlying model of codes, terms, concepts, and code attributes, illustrating the utility of metadata for coding systems and data standards. Metadata standards that can “wrap” all terminologies to some abstract features, so that they can be interchanged and related automatically by computers, will facilitate dealing with overlaps in standards.

Differences in terminological structures of candidate data sources influence both strategy and quality of mapping activities. Mapping is the deliberate act of determining equivalence (or acceptable measure of equivalence for a given context) of concepts from one terminology representation to another, and is an intentional non-trivial process that involves both domain knowledge and understanding of both terminology structures, and a clear understanding for
the intended uses that the mapping is to support. Though many view mapping across disparate terminologies as a solution for dealing with standards overlap, there are serious limitations in this approach. Mapping concepts from one terminology to the concepts in another is often not possible without losing data precision or intended semantics, especially when mapping between terminologies with varied levels of precision. The wide-ranging interests of the clinical research community might make it impossible to eliminate data standards overlaps, so strategies for integrating multiple (dynamic) data standards while maintaining data integrity will be a ripe area for future attention as will tools that facilitate this process. We expect that the NLM will coordinate this activity, but the clinical research community must define specific use cases and directions by which mappings should occur.

**Starting Points**

There are many opportunities for moving toward the use of data standards in clinical research. This identification of gaps and overlaps can be a starting point, but we expect that the conceptualization and scope of the standards areas presented in Table 1 can be expanded and refined. We hope that this summary of clinical research data standards will stimulate focused discussion on why, where, and how to achieve data standardization in this domain. Clinical research is at the leading and changing edge of variable invention, and many clinical research observations and measures are in such a state of flux that it may not be possible or important to standardize them. A nationally-driven clinical research data standards task force could illuminate which areas are priorities and develop informed and representative teams for strategically achieving useful and viable data standards in those areas.

Clearly, there is an overlap among many research and clinical variables, including laboratory, physiologic, and patient assessment measurements, which should be the area of first focus. The successful adoption of data standards (harmonized between clinical and research domains) for these areas will likely be contingent upon strong use cases that demonstrate benefits of shared standards. Best interoperability outcomes will result when the same terminologies are used in these areas, and it is imperative that the terminology models implemented in both domains are the same. The current course of developing information models and later adding terminologies for plug-and-play carries enormous risks of creating information silos within clinical research applications and between clinical research and clinical care activities. Active and early dialog between both communities at the time of development, and a dedication to using same terminologies in the same way, will enable harmonized standards within and across these communities.

Most of the research specific gaps are being addressed by CDISC. Since that group has strong industry participation, it is well-suited to take the lead. The participation of NIH, whose interests go beyond regulated clinical research, can ensure that needs are met through robust new standards that address broader clinical research interests. If CDISC functions as a workgroup to inform HL7 and ballots via established HL7 consensus procedures, synchronization of clinical and research interests will be likely.

Efforts that reduce data variation at the point of collection will simplify standardization processes in the future. Continuous variables (e.g., 25 cigarettes/day) should be collected instead of categorical variables (e.g., “smokes 0-1 packs/day”) whenever possible to allow future sharing and aggregation of data. The establishment of standards for question modeling (e.g., semantics in the question: Q: “Wheezing present?” A: “Yes/no”; semantics in the answer: Q: “Findings?” A: “Wheezing”; or a combination of both: Q: “Abnormal respiratory findings?” A: “Wheezing”) can practically eliminate significant differences in the representation and transmission of clinical data variables—in research and health delivery applications.

Much research data collection is done via forms, much of which has the look of a survey instrument and could be conceptualized as such. Activities that encourage the use of standardized questions and case report forms at the time of data collection will be valuable. Opportunities for investigators to share questions administered from data collection forms or standardized instruments are limited by their ability to understand and access the content of questions previously used by themselves or other investigators. Successful management of questions on existing data collection forms will support the re-use of existing items and their relevant coding into appropriate standardized terminologies. Addressing this much-needed ability to understand and access the content of standardized questionnaires could also increase the use of standards, and reduce the time that new investigators spend generating new question content.

**Conclusions**

Dependencies between health care data and research data are unavoidable, suggesting that clinical researchers should seek opportunities to share health care data standards where possible. Although data constructs unique to clinical research tend have gaps, there is significant overlap in the types of data that are captured for both clinical research and patient care—e.g., signs and symptoms, findings, observations, procedures, patient outcomes. It seems reasonable that the same terminological data standards would be used for both healthcare delivery and clinical care. Similarly, the harmonization of clinical care and research requires compatible information models and clinical researchers should remain abreast of (and participate in) development in standards for clinical care data and systems. Coordination is required to ensure that standardization movements in both the health care and the clinical research domains evolve in tandem.

Data standards are the critical foundation of the proposed national health information infrastructure. The importance of clinical research data within this infrastructure is underscored by new emphasis on translational science goals. The current strategy of federal standards efforts is to create an “interlocking set” of data standards for all of healthcare. An assessment of available standards is a prerequisite for understanding how the “pieces” (i.e., candidate data standards) are to be assembled, and only a survey of the clinical research domain and its unique requirements can measure whether the resulting “set” of data standards is suitable for the data representation purposes of clinical research. The gaps and overlaps of data standards for clinical research
data need to be resolved, which will take cooperation across the broad spectrum of clinical research interests. The complexity of choosing common data standards for clinical research arises not only from the number and diversity of interests in the clinical research community, but also from technical issues related to the structures and intended uses of various candidate data and terminological standards. The co-evolution of technology, definition of clinical research requirements, and the definition of data standards in health care delivery could result in common standards and applications demonstrating their utility. It is hoped that this early characterization of clinical research data constructs and current standards coverage will help focus an agenda for clinical research data standards and encourage discussion between clinical research and informatics communities.

References


