

CDE Development Model for Chest CT Screening for Lung Cancer

Curtis Langlotz, MD, PhD

Background

Several years ago, the NCI embarked on an effort to standardize data collection methods across the cooperative trials groups it funds. These data collection methods included common toxicity criteria (CTC) and common data elements, or CDEs. To date, CDEs have been developed for phase III therapeutic trials related to breast, prostate, lung, and colon cancer. Recently, the NCI funded the American College of Radiology Imaging Network (ACRIN, <http://www.acrin.org>), a cooperative group dedicated to imaging trials. Simultaneously, new imaging tests that have shown promise in screening for lung cancer. Thus, several groups within the NCI became keenly interested in developing CDEs for imaging trials. What follows is a description of the development of CDEs for chest CT screening for lung cancer that was undertaken in the last half of 1999.

Motivations

Because there are a large number of existing medical terminological resources and standards, our first step was to determine whether others have already accomplished the same task or a related one. Terminological resources such as the Unified Medical Language System (UMLS, <http://www.nlm.nih.gov/research/umls/>) [1] and SNOMED International (<http://www.snomed.org>) already provide a rich vocabulary for encoding clinical data. In many cases, these resources can be adopted directly for use in data collection for clinical trials. A number of imaging-specific classification systems are also available. For example, we considered the American College of Radiology's Index for Radiological Diagnoses [2], a bi-axial classification that was originally developed for the organization of teaching files according to anatomical and pathological information. We also evaluated an extensive classification system for the indexing of scientific literature developed by the Radiological Society of North America [3] and a glossary of chest imaging published by the Fleischner Society [4].

Unfortunately, the results of our tests were in accord with the results of previous studies of existing terminological resources, which suggest that existing terminologies are lacking in many of the types of terms used in imaging. For example, while these terminologies contain a rich anatomic and pathological vocabulary, they rarely contain the descriptive visual features that radiologists use to describe findings on imaging studies or the technical terms that describe how

images were acquired and viewed. One study of the UMLS, SNOMED, and the ACR Index found that they contained less than half of terms used in ultrasound reporting [5]. Because we found similar results in a pilot study of chest CT terms (unpublished data), we embarked on the development of a comprehensive set of common data elements for chest CT screening, while relying on the UMLS and SNOMED to guide our use of anatomic and pathological terms.

Attributes of a Common Data Element (CDE)

The goal of the CDE development process was to produce CDEs that could be made available in a web-accessible database on the NCI's web site (<http://cii.nci.nih.gov/cde>). The database contains a searchable list of terms and their definitions, following a dictionary format. In this section we provide definitions of CDE attributes used in that database, and describe the hierarchy of CDE categories that was developed for chest CT screening CDEs. The following attributes are typically specified for each CDE.

1. *Category* is a grouping or classification into which a CDE falls. For example the Imaging CDE category is broken down into sub-categories, such as anatomic locations and imaging findings.
2. *Short CDE Term* lists a brief unique name for the CDE. For example “Exam Quality” is an imaging CDE that allows a radiologist or other imaging professional to rate the quality of the images being interpreted.
3. *Long CDE Term* indicates how a CDE might be listed on a data collection form. For example, “The overall diagnostic quality of the imaging study” is the Long CDE Term for the “Exam Quality” CDE.
4. *CDE Values* itemize the list of possible values that a CDE can take on. For example, the Exam Quality CDE can take on values of Optimal, Diagnostic, Limited, Non-diagnostic, and Uninterpretable.
5. *Definition*—a text description of a CDE Term or CDE Value. For example, the definition of the “Uninterpretable” CDE Value is “No useful diagnostic information. The study should be repeated”.

Because our CDE development effort was the first to focus on imaging trials, no groupings of imaging CDEs had been developed. We also found that no detailed imaging data model was available from other terminologies or standards.

Therefore, we drafted a hierarchy containing several new imaging CDE subcategories, which served as a simple data model (see Table 1). This hierarchy will be refined and augmented as additional imaging CDEs are developed.

Table 1: The CDE Categories developed for CT Imaging for screening of lung cancer.

*Imaging**Study Technique**Equipment**Acquisition**Protocol**Contrast Agent**Display**Exam Quality**Image Location**Anatomic Location**Lung Lobes**Right Lung**Left Lung**Mediastinum**Thoracic Lymph**Pleura**Metastases**Findings**Nodules**Other Findings**Conclusion**Recommendation***The CDE Development Process**

The CDE development process resulted in a comprehensive set of approximately 120 data elements over a period of 2-3 months. The process was conducted in cooperation with professional organizations (e.g. the Radiological Society of North America), standards development organizations (e.g., DICOM--Digital Imaging and Communications in Medicine), and interested scientific groups within the NCI. We followed the technical desiderata defined by Cimino et al [6] and the conceptual framework outlined by Campbell et al [7]. Because the technical and philosophical issues have been reviewed in this prior literature, we will focus here on the social, political, and administrative issues we faced. The steps we took in developing the CDEs for chest CT screening are detailed below.

CDE Development Steps

1. *Assemble pertinent resources.* The first step was to assemble a comprehensive set of resources from which proposed CDEs could be extracted. Even when some terminological resources are available in a given domain, there often are additional resources that can be helpful in generating a more comprehensive

Preliminary Version, Not for Distribution

- set of CDEs. These resources include schemas for clinical trial databases, data collection forms from ongoing clinical trials, publications of related preliminary scientific research (e.g., [8]), and portions of relevant standards and terminologies.
2. *Involve relevant stakeholders.* As investigators in ongoing clinical trials are contacted to obtain the resources above, they should be asked to participate in the CDE development process, either by attending panel meetings, participating in electronic mail discussions, or both. Our panel included predominantly clinical researchers and clinical experts—especially those with an interest in informatics and medical terminology. But we also tried to include other relevant stakeholders, including members of cooperative groups, biostatisticians, epidemiologists, and data coordinators. Leaders from professional organizations and standards development organizations nominated additional members to the panel. This resulted in a productive mix of expertise, interest, institutional support, and opinion leadership.
 3. *Create initial CDEs and CDE categories.* We convened a small working group of 3-5 people, consisting primarily of informatics and clinical experts to sort through the pertinent resources and create a preliminary list of data elements. In some cases, multiple sources provided similar or duplicate data elements. The working group organized the list of proposed elements into categories and subcategories, and created a reduced list of candidate CDEs to be considered by the CDE panel. When there appeared to be more than one appropriate method to represent a given data element, both possibilities were presented to the panel.
 4. *Link proposed CDEs to existing standards and terminologies.* Prior to the panel's deliberations, consistency between the proposed CDEs and existing standards and vocabularies was assessed. We felt that there might be important reasons for any differences between newer data collection methods and existing terminologies and standards. Consequently, prior to the panel meeting we weeded out potential conflicts only when we could identify clearly preferable existing approaches, rather than enforcing rigid consistency to existing terminologies in every case.
 5. *Distribute draft CDEs to the full panel.* Prior to the panel meeting, the panelists were sent a list of these draft CDEs in the form of a spreadsheet. Each panelist was encouraged to review the spreadsheet to formulate questions and comments that they felt should be discussed by the group.
 6. *Convene CDE panel.* Although a face-to-face panel meeting may not be necessary for every terminological development effort, we felt that the relative paucity of existing terminological efforts in imaging, together with the diverse backgrounds of the panelists, made an in-person meeting beneficial. The meeting allowed each member of the panel to be oriented to the concerns,

- needs, and skills of the others. A brief presentation of the results of a recent trial, with special emphasis on data collection methods, highlighted the clinical needs. A presentation of a tentative protocol design for an upcoming trial highlighted the potential near-term utility of CDEs. A demonstration of the CDE web site provided an informatics emphasis and a focus on the overall goal of the process. Following this brief orientation, the panel deliberated for about 5 hours.
7. *Finalize CDEs.* Following the panel's deliberations, the CDEs were revised by the informatics working group to reflect the discussion. Additional comments on the revised version were sought via electronic mail. The relationship to other terminologies and standards were rechecked using the UMLS knowledge sources [1] and the published DICOM standard [9].
 8. *Publish CDEs in web-accessible database.* After the panel finalized the new CDEs, the list was provided in spreadsheet form to the NCI's Office of Informatics for inclusion in the web-accessible CDE database. We then sought to publicize the availability of the new CDEs among the communities of clinical researchers that may find them useful.

Avoiding Pitfalls of CDE Deliberations

To prevent unnecessarily lengthy discussions of side issues, we laid out several principles to guide the panel's deliberations. All of us naturally become attached to the terms we use in our daily work. Likewise, cooperative groups and other large research organizations typically have invested considerable resources in their data collection methods. These attachments, together with the heterogeneity of existing data collection systems creates the potential for lengthy discussions regarding the choice between forms of a common data elements. We therefore outlined the following principles for the panel prior to its deliberations:

- 1) *Clinical research, not clinical practice.* The discussion should focus on common data collection methods for cancer research, not for clinical practice. For example, we focused on standard methods to collect data about how chest CT screening studies were performed. Any discussion of standards for how chest CT screening studies *should* be performed clinically was considered outside our purview, since attempts to formulate such standards by the NCI would be viewed with skepticism by professional organizations, who view the creation of such standards one of their key roles.
- 2) *Data collection, not experimental design.* The panel should address how data should be collected for clinical trials, rather than how those trials should be designed and conducted. For example, the group did not discuss the merits of whether a particular data item should be used as an eligibility criterion or a stratification variable. Instead, the panel focused on how that data item should

- be collected and encoded if trial designers elect to collect it as part of a clinical trial.
- 3) *Adoption rather than re-invention.* The panel should be strongly predisposed toward adopting previous terminological standards when possible, rather than “reinventing the wheel”. This principle applies not only to existing standards and terminologies, but also to commonly-used anatomic classification and staging systems. For example, rather than reinventing a staging system for classifying mediastinal lymph nodes, a standard, widely-used classification system [10] was adopted.
 - 4) *Synonyms, not competitors.* Occasionally, two groups use two different terms to describe essentially the same phenomenon. In that case, informatics systems can consider these two terms as synonyms [6], so that each group can use the familiar term to describe the same concept. If each synonym has the same definition, the two synonyms will be used consistently. Thus, terminological “synonym wars” can be avoided. On the other hand, there are occasional semantic differences between similar terms that should be discussed openly and resolved. If the differences cannot be resolved, the two terms should both be retained, each with a detailed definition and specification of the difference between the similar terms.
 - 5) *Expertise, not turf.* The interdisciplinary panel generally should defer to the individual with the training and experience most specific to the term under consideration. For example, most imaging trials will collect extensive information about gross pathologic specimens, in order to maintain a reference standard. Although most of the panel members may be radiologists, a pathologist or surgeon is probably best able to resolve terminological issues related to gross pathologic specimens.

Discussion

Using the procedures described above, we created a set of about 120 common data elements (CDEs) over a period of approximately three months. The CDEs were created in cooperation with the several groups within the National Cancer Institute, including the Biomedical Imaging Program, the Lung and Aerodigestive Cancer Research Group, the Office of Informatics, and the SPORE Program. Other cooperating organizations included the American College of Radiology Imaging Network (ACRIN), the Radiological Society of North America (RSNA), and the Digital Imaging and Communication (DICOM) Working Group 18 on Clinical Trials and Education. Since ongoing CDE development has ongoing importance to NCI’s mission, we will now consider several issues related to NCI’s continuing role in developing terminology.

It may be beneficial for NCI to strengthen its relationship to other standards-development organizations, such as HL7, SNOMED, and DICOM. For example,

the NCI could actively monitor and contribute to ongoing these standards activities. This will contribute to the NCI's ability to rapidly develop CDEs that contribute to a convergent and comprehensive medical terminology.

As clinical trial sites adopt the CDEs, the NCI's role will naturally shift from the process of developing CDEs and the methods for their use, to the process of encouraging ongoing use and managing change. NCI's role at that point will shift to measuring CDE usage, providing technical assistance to support adoption and implementation, and supporting a change-management process that involves all stakeholders.

There are also several technical enhancements for NCI to consider. For example, the CDEs have been viewed essentially as dictionaries, consisting of a linear list of terms, each with its own definition. However, most successful terminologies have developed a rich set of knowledge about each term that not only provides linkages between terms and a logical context for terms, but also specifies how a term should be instantiated and used in a variety of contexts. The NCI's CDE "Categories" can be viewed as a nascent information model for CDEs. With time, the CDE Categories must evolve into a richer representation of the relationships among terms and term groups. Ultimately, sophisticated knowledge representation techniques, such as semantic networks and description logic, will be required to express adequately the relationships among terms and to facilitate convergence with existing information models. Likewise, more detailed data dictionary entries for new data elements will minimize variation in term meaning and usage according to context and user. In some cases, imaging information or other non-textual data should augment simple text descriptions.

The current web-enabled interface to the CDE database is designed to encourage consistent data collection not only from the cooperative groups, but also from oncology trials being conducted in other settings. The ability to perform boolean keyword searches in a transparent user interface, and the ability to display and download terms in forms other than HTML will likely enhance its usability to the disparate constituencies that it is intended to serve. For example, users without a great deal of research infrastructure may wish to download CDEs as draft data-collection forms or draft database specifications.

Conclusion

We have described how a well-defined set of common data elements can be created in a relatively short period of time. We have focused primarily on the social, political, and administrative issues, since others have comprehensively discussed the technical and conceptual issues [6, 7]. The resulting CDEs are currently available on the NCI's web site, and likely will become a part of both SNOMED and the UMLS. The results have been officially sanctioned by the board of directors of the Radiological Society of North America, and will be

employed in the design of an upcoming chest CT screening trial to be supported by the National Cancer Institute. We have described this process in detail to allow others to learn from our experience and avoid some of the pitfalls we identified, thereby resulting in more efficient terminological development efforts. We also hope our efforts will identify key informatics research issues that will be the underpinnings of a rich set of informatics resources for clinical research and clinical care.

Acknowledgments

I would like to thank a number of people whose efforts were vital to this manuscript and to the CDE development process it describes. Dan Sullivan, who directs the Biomedical Imaging Program at the National Cancer Institute, supported efforts to develop imaging CDEs from the beginning. John Silva provided the leadership that made CDEs a reality. Jeff Abrams and Beverly Meadows shared their valuable experience with CDE development for therapeutic trials. Christine Berg and Jorge Gomez from the Division of Cancer Prevention of the National Cancer Institute provided support for the CDE panel meetings. Denise Warzel and Carolyn Pifer were close collaborators on the creation of draft CDE documents. Marie Zininger from the American College of Radiology and Dana Davis from the Radiological Society of North America facilitated the involvement of professional societies. This manuscript began as an appendix to the report of the NCI's Informatics Long Range Planning Committee. I am grateful to the members of that committee, who provided comments on earlier drafts. Finally, I would like to thank Marion Ball and Judy Douglas, whose able leadership and gentle encouragement made this chapter (and this book) a reality.

References

1. National Library of Medicine, *Unified Medical Language System Knowledge Sources*. 10th ed. 1999, Bethesda, MD: U.S. Department of Health and Human Services. 147.
2. American College of Radiology, *Index for Radiological Diagnoses*. 4th ed. 1992, Reston, VA: ACR.
3. Radiological Society of North America, *RSNA Index to Imaging Literature*. 1999, Oak Brook, IL: Radiological Society of North America.
4. Tuddenham, W. (1984). Glossary of terms for thoracic radiology: Recommendations of the Nomenclature Committee of the Fleischner Society. *AJR* 143:509-517.
5. Bell, D. and Greenes, R. (1994). Evaluation of UltraSTAR: Performance of a collaborative structured data entry system. *JAMIA Symposium Supplement*:216-222.

6. Cimino, J., Clayton, P., Hripcsak, G., *et al.* (1994). Knowledge-based approaches to the maintenance of a large controlled medical terminology. *JAMIA* 1(1):35-50.
7. Campbell, K., Oliver, D., Spackman, K., *et al.* (1998). Representing thoughts, words, and things in the UMLS. *J Am Med Inform Assoc* 5(5):421-431.
8. Henschke, C., McCauley DI, Yankelevitz, D., *et al.* (1999). Early Lung Cancer Action Project: overall design and findings from baseline screening. *Lancet* 354(9173):99-105.
9. Digital Imaging and Communications in Medicine, *DICOM Information Object Definitions*. 1999, Rosslyn, VA: National Electrical Manufacturers Association.
10. Mountain CF and Dresler, C. (1997). Regional lymph node classification for lung cancer staging. *Chest* 111(6):1718-1723.