CIN PLUS

Section Editor: Mollie R. Cummins, PhD, RN, FAAN

Case Study

Semantic Annotation of a Pediatric Critical Care Research Study

Katherine A. Sward, PhD, RN, Sarah Rubin, MD, Tammara L. Jenkins, MSN, RN,
Christopher J. Newth, MD, FRCPC, J. Michael Dean, MD, for The Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) Collaborative Pediatric Critical Care Research Network (CPCCRN)

Clinical care, research, and quality initiatives such as the Learning Health System require organizations to share and understand each other's data.[1] Such interoperability requires standard messaging formats and standard terminologies.[1,2] The Office of the National Coordinator for Health IT seeks to achieve electronic health record (EHR) connectivity before 2018, but some believe that, despite limited successes, it will take at least another decade before interoperability is realized on a national scale.[3] Establishing common understanding across all stakeholders is complicated, in part because the meaning (semantics) and format of terms are often context dependent. Thus, it remains important to assess the extent to which standard terminologies appropriately represent clinical meaning in specific contexts. This article reports a case study in which a pediatric critical care research network registry was mapped to standard terminologies. The University of Utah Institutional Review Board approved this study; this evaluation did not involve human subjects or patient data.

KEY POINTS:

- Interoperability remains an important national informatics focus
- Nuances of meaning can be lost during terminology efforts
- It is important for nurses with clinical expertise to be aware of and engage in interoperability related initiatives

STANDARD TERMINOLOGIES

Standard terminologies represent concepts (ideas or things), identified with a code and a human-readable name. Definitions and synonyms may also be included. There are many standard terminology systems, of which five (Systematized Nomenclature of Medicine Clinical Terms [SNOMED-CT], Logical Observation Identifiers Names and Codes [LOINC], RxNorm, Current Procedure Terminology 4 (CPT-4), and International Classification of Diseases, 10th Edition [ICD-10]) have been mandated for use in EHRs as part of the Affordable Care Act. The National Library of Medicine coordinates across these and other American Nurses Association-recognized standard terminologies, within the Unified Medical Language System (UMLS) terminology services.[4] Pediatric-specific terms have been generally lacking in standard terminologies.[1,5,6] The *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) led a harmonization effort focused on improving the coverage of pediatric terms. Initial efforts described child life stages and terms from neonatal research. Pediatric critical care terms were not the focus of initial formal harmonization efforts.[1] However, terms in pediatric critical care may differ in nuanced ways from the way terms are defined in other areas. Ensuring that datasets are defined in ways that are relevant to pediatric critical care is crucial if evidence derived from such endeavors is to be clinically relevant in that context.

The harmonized NICHD Pediatric Terminology is housed within the National Cancer Institute (NCI) standard terminology system, called Enterprise Vocabulary Services (EVS). The EVS was designed to support many types of research. Basic concepts can be combined to represent more complex ideas.[7] Concepts are identified with an alphanumeric code (eg, C16696) and a definition that helps differentiate terms like *discharge*, when meaning leaving the institution versus when the word means wound drainage. Although humans are very good at such disentanglement, computers need each meaning to be explicitly differentiated.

HISTORY OF THE CASE STUDY

The Collaborative Pediatric Critical Care Research Network (CPCCRN) is a network of pediatric intensive care units (PICUs) with centers across the United States.[8] During 2012 to 2013, the CPCCRN conducted a study examining a

Author Affiliations: College of Nursing (Dr Sward) and Department of Biomedical Informatics, School of Medicine (Drs Sward and Dean), University of Utah, Salt Lake City, Utah; Department of Anesthesiology and Critical Care Medicine, Children's Hospital Los Angeles (Drs Rubin and Newth); Keck School of Medicine, University of Southern California, Los Angeles, California (Drs Rubin and Newth); National Institute of Child Health and Human Development (NICHD), Bethesda, Maryland (Jenkins); and Department of Pediatrics, Division of Pediatric Critical Care, School of Medicine, University of Utah, Salt Lake City, Utah (Dr Dean).

Supported, in part, by the following cooperative agreements from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Department of Health and Human Services: U10HD050096, U10HD049981, U10HD049983, U10HD050012, U10HD063108, U10HD063106, U10HD063114, and U01HD049934.

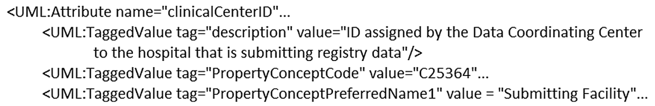
Sarah Rubin is a KL2 Scholar awarded under the KL2 Mentoring Research Career Development Award (National Center for Advancing Translational Science, National Institutes of Health) through Southern California Clinical and Translational Science Institute at University of Southern California, Keck School of Medicine.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The authors have disclosed that they have no significant relationship with, or financial interest in, any commercial companies pertaining to this article.

Corresponding author: Katherine A. Sward, PhD, RN, University of Utah, Salt Lake City, UT 84112 (kathy.sward@nurs.utah.edu)

Copyright © 2016 Wolters Kluwer Health, Inc. All rights reserved.



```
<UML:Attribute name="clinicalCenterID"...
      <UML:TaggedValue tag="description" value="ID assigned by the Data Coordinating Center
              to the hospital that is submitting registry data"/>
      <UML:TaggedValue tag="PropertyConceptCode" value="C25364"...
      <UML:TaggedValue tag="PropertyConceptPreferredName1" value = "Submitting Facility"...
```

FIGURE 1. Example of semantic annotation tags within an XML file.

potential data sharing mechanism, using NCI open-source tools to create an interoperability infrastructure called picuGrid.[8] The picuGrid was based on an ongoing registry study that uses data from hospital administrative databases to describe the population of the PICUs. Automatically extracting and integrating clinically collected data have been proposed as a way to make conducting multicenter studies more efficient.[6,9] Developing the picuGrid included a semantic annotation step that allowed us to harmonize with the NICHD Pediatric Terminology efforts.

SEMANTIC ANNOTATION

Semantic annotation is a process of attaching names, codes, descriptions, or other metadata to a model or text document. We created a Unified Modeling Language information model.[2,10,11] An information model defines common data elements (CDE), with the category (such as patient), data element name (eg, birthdate), definition, and the set of possible responses.[12] An example is Patient.DateOfBirth, defined as "The calendar date on which a person was born," with an actual date as the allowable response. The model can be displayed graphically and can be easily exported as an XML file.

The NCI-developed semantic tools (Semantic Integration Workbench) automatically generate a list of concepts from the EVS terminology that potentially match each CDE in the information model. Because the NICHD pediatric terminology is housed within the EVS, we could preferentially select concepts from the NICHD pediatric terminology, when one was available. A nurse with critical care experience identified the applicable matches. The NCI tools then embedded tags into the information model representing the selected EVS concepts; this process is called semantic annotation. A segment of XML with annotation tags can be seen in Figure 1.

EVALUATION METHODS

Annotations were verified by the EVS team using their standard verification processes and the picuGrid annotated information model was published in the EVS as a use-case (subset) of the NICHD pediatric terminology.[1,2,10,11] In addition, we manually mapped the data elements to other terminology systems (Table 1) and conducted a descriptive gap analysis to determine the coverage within each terminology system.

RESULTS

We defined main concepts as the items in the CPCCRN registry study protocol. The picuGrid data model had 21 main concepts including admission and discharge dates, patient demographics, and lists of diagnoses and procedures. The information model also included three items we considered as record-keeping concepts, such as a timestamp (items relevant to the database but not part of the research protocol). The model contained 30 enumerated values. For example, one main concept was payer type; the enumerated values were responses like Medicaid, insurance, and self-pay.

Table 1. Terminology Systems Evaluated in This Study

Name—Organization	Description/Highlights	URL
EVS—NCI[2,7,8,12]	Houses the NICHD pediatric terminology; contains CDEs with explicit definitions	http://evs.nci.nih.gov
Logical Observation Identifiers Names and Codes (LOINC)—Regenstrief Institute[a,13]	Standard terminology identifying laboratory tests, observations, and assessments	https://loinc.org
PhenX Toolkit—RTI International; funded by the National Human Genome Research Institute (NHGRI)[14]	Standard measures for complex diseases, phenotypic traits, and environmental exposures; embedded within LOINC	https://www.phenxtoolkit.org
UMLS—National Library of Medicine[15]	Many terminology, classification, and coding standards; includes access to SNOMED-CT. The EVS terminology was initially derived as a subset of the UMLS.	http://www.nlm.nih.gov/research/umls
SNOMED-CT— International Health Terminology Standards Development Organization (IHTSDO)[a,15]	A comprehensive clinical terminology that is internationally recognized and widely used for clinical data	http://www.ihtsdo.org/snomed-ct
ICD—World Health Organization[a,16]	Used to categorize causes of mortality and morbidity and to monitor the incidence and prevalence of diseases. Used to code data for billing and reimbursement in the United States	http://www.who.int/classifications/icd/en

[a] Standards designated by the US Federal Government for the electronic exchange of clinical health information.[17]

Copyright © 2016 Wolters Kluwer Health, Inc. All rights reserved.

GAP ANALYSIS

All of the items in the picuGrid model were able to be represented by concepts in the EVS terminology. For the main picuGrid concepts (items listed in the protocol for the CPCCRN registry study), six of the 21 concepts (29%) were found in the NICHD pediatric terminology, the remainder were found elsewhere within the EVS. A single EVS code was sufficient to represent 21% of the concepts and 70% of the enumerated values (overall, 50% of the picuGrid data model). Single EVS codes (precoordinated terms) represented demographic variables like patient age, but a combination of codes (postcoordination) was needed to appropriately represent other information. For example, hospital admission date required three codes: hospital (C16696) + admission event (C25385) + date (C25164). Codes had to be combined to represent Diagnosis-Related Group (DRG) and Major Diagnostic Category (MDC). At the time of this study, the International Classification of Diseases, Ninth Edition (ICD-9) was in use rather than ICD-10. There was a single EVS code for ICD-9, but this had to be combined with other codes to designate subsets (diagnosis ICD-9, procedure ICD-9, and external cause of injury [e-code]). Pediatric intensive care unit (ICU) had to be constructed by combining the codes for pediatric (C39299) and the code for ICU (C53511). Table 2 shows the number of codes that were needed to identify the picuGrid data elements.

ADDITIONAL ANALYSES

We manually mapped to other terminologies (Table 1). We found that 84% of the picuGrid concepts were represented by LOINC, which are required by meaningful use regulations to represent observations. Unlike the EVS, LOINC does not allow codes to be combined. Other notable differences between LOINC and the EVS were also seen. Instead of PICU admission date and PICU discharge date, LOINC contained a code for days in the ICU. Hospital admission date and hospital discharge date, however, are separate LOINC codes. There was a LOINC code for ICD-9, but like EVS, there were no ICD-9 subsets. There was a code for newborn ICU in LOINC, but not a code for pediatric ICU. We also examined PhenX, an emerging terminology that represents genetic and phenotypic information that is embedded within LOINC. Only three of our data elements were found in PhenX (gender, race, and ethnicity). Finally, we mapped picuGrid CDEs to the UMLS, in part because the EVS was initially derived from the UMLS, and because the UMLS integrates a large number of standard terminology systems including SNOMED and nursing terminologies. All picuGrid concepts were found in the UMLS. There were single codes in the UMLS for some items that had to be constructed from multiple codes in the EVS terminology. Most notably, there were single codes for DRG (C0011928), MDC (C1550395), and PICU (C0021710). Like other terminologies, the UMLS did not have distinct codes for ICD-9 subsets.

Table 2. Number of Codes to Map a CPCCRN Registry Data Element to the EVS Terminology

Registry Data Model Category	Registry Data Element	Number of EVS Codes Needed	NICHD Pediatric Terminology Category
Patient	Patient ID	2	Demographics
	Date of birth	1	Demographics
	Gender	1	Physical exam
	Race	1	Demographics
	Ethnicity	1	Demographics
Hospital visit	Clinical center ID	2	
	Hospital admission date	3	
	Admission type (emergency/ scheduled)	2	
	Zip code for patient at time of visit	1	Demographics
	Primary payer	1	
	DRG code	4	
	DRG text	4	
	MDC code	4	
	MDC text	4	
	Hospital Discharge date	3	
	Discharge Disposition	2	
PICU stay	PICU admission date	4	
	PICU discharge date	4	
Diagnosis	Diagnosis codes (ICD-9)	3	
Procedure	Procedure codes (ICD-9)	3	
eCode	e-Codes (ICD-9)	4	
(record keeping)	Data Send Indicator	3	
	Row Identifier	1	
	Timestamp	2	

e-Codes refers to ICD-9 external cause of injury.

DISCUSSION

We found that common administrative PICU data elements were well represented in standard terminology systems. The EVS terminology was able to represent all of the picuGrid concepts, but only demographic items matched the NICHD pediatric terminology. That appeared to be primarily because (1) the items in our study were administrative data not unique to pediatrics, and (2) we used broad category names (e.g., diagnosis ICD-9 code) whereas the NICHD Pediatric

Copyright © 2016 Wolters Kluwer Health, Inc. All rights reserved.

Terminology lists specific diagnoses by name. As has been found in other clinical domains, representation of pediatric critical care concepts required a combination of multiple codes (post-coordination). Combined codes may provide a literal definition of the name and were adequate for our picuGrid study because there was common understanding of meaning across sites, based on the registry study protocol. However, postcoordinated terms may not fully express nuances of meaning[13] and may not have the same meaning for all stakeholders.

In most of the evaluated terminologies, codes had to be combined to represent pediatric intensive care; however, many of the terminologies did have precoordinated codes for neonatal intensive care. This may be because pediatric intensive care is a slightly "newer" specialty area than neonatal intensive care. It may also be pragmatic; terms are added to the EVS (and some of the other terminology systems) when they are requested by investigators, and the neonatal ICU was part of the NICHD initial harmonization project.

In addition, DRG and MDC had to be constructed from multiple codes. This was a surprising finding. Codes for DRG and MDC are used for reimbursement, which is a specific, nuanced financial meaning, so we expected that these would be represented as a single precoordinated code. We needed to use multiple codes to represent ICD subsets, also. This was less concerning because those subsets are distinguished internally within ICD, which is a standardized terminology in and of itself.

Our study had limitations. The picuGrid model was based on a registry study that describes the PICU population in general. Many clinical studies include assessments or procedures that may be unique to the domain; we may have found fewer matches had we chosen a clinical study for the exemplar.[1] We did not separately identify nursing-specific data elements; such efforts are increasingly important to show nursing's contribution to sharable and comparable data.[4]

Our study also had strengths. We used well-established techniques and NIH-developed tools. The study team included experts in nursing informatics, computer science, and pediatric critical care. Data elements were agreed upon by multiple centers, and we prioritized compatibility with national NICHD terminology efforts, so our results should be at least somewhat generalizable. The main analysis corresponded well with examination of other national terminologies.

IMPLICATIONS FOR NURSING INFORMATICS PRACTICE

This case study demonstrates the importance of continuing efforts to examine nuances of interoperability. National initiatives to create sharable, comparable nursing data are ongoing and have been a focus of the Nursing Knowledge/Big Data Science meetings at the University of Minnesota. It is crucial for nurses with clinical expertise to be aware of, and engage in, these terminology efforts to ensure that nuances of meaning are not lost.

Acknowledgments

Thanks go to Lewis J. Frey, for instrumental leadership on the picuGrid project. Grateful appreciation is also expressed to the CPCCRN investigators, who provided valuable insights and feedback and without whom the study could not have been completed, including Kathleen L. Meert, Robert A. Berg, Joseph A. Carcillo, Murray M. Pollack, Thomas P. Shanley, Heidi J. Dalton, David L. Wessel, Rick E. Harrison, Allan Doctor, and Robert Tamburro.

References

1. Forrest CB, Margolis PA, Bailey LC, et al. pedsnet: a national pediatric learning health system. *J Am Med Inform Assoc*. 2014;21: 602–606. doi:10.1136/amiajnl-2014-002743.
2. Min H, Ohira R, Collins MA, et al. Sharing behavioral data through a grid infrastructure using data standards. *J Am Med Inform Assoc*. 2014;21: 642–649. doi:10.1136/amiajnl-2013-001763.
3. Pallardy C. Gorilla CIO: 'the industry will be cracking health IT interoperability for the next decade'. *Becker's Health IT and CIO Review 2016*. 2016.
4. Warren JJ, Matney SA, Foster ED, Auld VA, Roy SL. Toward interoperability: a new resource to support nursing terminology standards. *Comput Inform Nurs*. 2015;33(12): 515–519. doi:10.1097/CIN.0000000000000210.
5. Hirschfeld S, Samavedam R, Keller M, et al. *Advancing child health research through harmonized pediatric terminology.* San Francisco, CA: Paper presented at: AMIA Summit on Clinical Research Informatics; 2010.
6. Hutton JJ. *Pediatric Biomedical Informatics: Computer Applications in Pediatric Research*. London, England: Springer; 2012. doi:10.1007/978-94-007-5149-1.
7. de Coronado S, Wright LW, Fragoso G, et al. The NCI Thesaurus quality assurance life cycle. *J Biomed Inform*. 2009;42(3): 530–539. doi:10.1016/j.jbi.2009.01.003.
8. Frey LJ, Sward KA, Newth CJ, et al. Virtualization of open-source secure web services to support data exchange in a pediatric critical care research network. *J Am Med Inform Assoc*. 2015;22(6): 1271–1276. doi:10.1093/jamia/ocv009.
9. Buetow KH. An infrastructure for interconnecting research institutions. *Drug Discovery Today*. 2009;14(11-12): 605–610. doi:10.1016/j.drudis.2009.03.011.
10. Kunz I, Lin MC, Frey L. Metadata mapping and reuse in caBIG. *BMC Bioinformatics*. 2009;10(suppl 2): S4. doi:10.1186/1471-2105-10-S2-S4.
11. Phillips J, Chilukuri R, Fragoso G, Warzel D, Covitz PA. The caCORE Software Development Kit: streamlining construction of interoperable biomedical information services. *BMC Med Inform Decis Mak*. 2006;6: 2.
12. National Library of Medicine. NIH Common Data Element (CDE) resource portal: glossary. 2013. http://www.nlm.nih.gov/cde/glossary.html. Accessed February 18, 2015.
13. Cooper M. National Cancer Institute caDSR best practice: pre-coordinated and post-coordinated terms. 2012. https://wiki.nci.nih.gov/display/caDSR/caDSR+Metadata+Development+Best+Practices;jsessionid=01B65E1F85F2C20270F0C860C5CA1981. Accessed January 28, 2016.
14. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX Toolkit: get the most from your measures. *Am J Epidemiol*. 2011;174(3): 253–60.
15. National Library of Medicine. Unified Medical Language System® (UMLS). 2014. http://www.nlm.nih.gov/research/umls/. Accessed February 20, 2015.
16. Caskey R, Zaman J, Nam H, et al. The transition to ICD-10-CM: challenges for pediatric practice. *Pediatrics*. 2014;134(1): 31–6.
17. Office of the National Coordinator for Health IT [ONC]. Standards and Certification Regulations. 2014. http://www.healthit.gov/policy-researchers-implementers/standards-and-certification-regulations. Accessed March 18, 2015.

Copyright © 2016 Wolters Kluwer Health, Inc. All rights reserved.