Model Formulation ■

A Model for Evaluating Interface Terminologies

S. Trent Rosenbloom, MD, MPH, Randolph A. Miller, MD, Kevin B. Johnson, MD, Peter L. Elkin, MD, Steven H. Brown, MD

Abstract **Objective:** Evaluations of individual terminology systems should be driven in part by the intended usages of such systems. Clinical interface terminologies support interactions between healthcare providers and computer-based applications. They aid practitioners in converting clinical "free text" thoughts into the structured, formal data representations used internally by application programs. Interface terminologies also serve the important role of presenting existing stored, encoded data to end users in human-understandable and actionable formats. The authors present a model for evaluating functional utility of interface terminologies based on these intended uses.

Design: Specific parameters defined in the manuscript comprise the metrics for the evaluation model.

Measurements: Parameters include concept accuracy, term expressivity, degree of semantic consistency for term construction and selection, adequacy of assertional knowledge supporting concepts, degree of complexity of pre-coordinated concepts, and the "human readability" of the terminology. The fundamental metric is how well the interface terminology performs in supporting correct, complete, and efficient data encoding or review by humans.

Results: Authors provide examples demonstrating performance of the proposed evaluation model in selected instances.

Conclusion: A formal evaluation model will permit investigators to evaluate interface terminologies using a consistent and principled approach. Terminology developers and evaluators can apply the proposed model to identify areas for improving interface terminologies.

■ **J Am Med Inform Assoc.** 2008;15:65–76. DOI 10.1197/jamia.M2506.

Introduction

Developers should create and evaluate clinical terminologies based on the terminology's intended usage.[1] The authors previously described interface terminologies as unique vehicles for supporting efficient and accurate interaction between healthcare providers and computer-based clinical applications.[2] Such applications often have difficult-to-use, internally structured data representations. Healthcare providers generally use interface terminologies to accomplish one of two tasks: 1) encoding clinical narrative into a structured form, or, 2) reviewing structured clinical information that has previously been encoded using a different terminology. In supporting such uses, interface terminologies must enable correct and rapid interaction between clinicians and structured clinical data, support facile use by healthcare providers through easy understandability, and integrate well with other clinical computerized systems in the environment (Rogers J, personal communication, 2003).[1–4] The manuscript presents an evaluation model defining and utilizing the attributes to measure interface terminology application system usability.

Background

A clinical terminology system consists of a collection of words or phrases organized together to represent the entities and relationships that characterize the knowledge within a given biomedical domain.[5–7] Each clinical terminology system meets distinct functional needs within a healthcare setting.[2,4,8–10] Terminology developers and users employ terminologies to represent the clinical knowledge of specific medical domains. Others use terminologies to classify clinical procedures for reimbursement purposes, or to categorize and index the content of scientific publications. Clinicians and researchers use other terminologies to characterize therapeutic uses and physiologic effects of medications. Increasingly, practitioners using electronic medical record systems employ specific terminologies to support generation of clinical documents.

Previously, terminology evaluators called for a single, uniform, standard approach to the creation, evaluation and adoption of all clinical terminologies.[6,7,11–17] Commonly employed early models for clinical terminology evaluation[6,16–20] sought to determine whether a terminology provided an exact and complete representation of a given

Affiliations of the authors: Department of Biomedical Informatics (STR, RAM, KBJ, SHB), School of Nursing (STR, RAM), Department of Pediatrics (STR, KBJ), Vanderbilt University, Nashville, TN; Mayo Foundation for Medical Education and Research (PLE), Rochester, MN; U.S. Department of Veterans Affairs (SHB), Nashville, TN.

This project was supported by Grants from the United States National Library of Medicine (Rosenbloom, 5K22 LM008576-02; Miller 5R01 LM007995) and from the Centers for Disease Control and Prevention (Elkin, PH000022-02 and HK00014-01).

Correspondence: S. Trent Rosenbloom, MD, MPH, Eskind Biomedical Library, Room 440, 2209 Garland Avenue, Nashville, TN 37232-8340; e-mail: <trent.rosenbloom@vanderbilt.edu>.

Received for review: 05/09/07; accepted for publication: 08/08/07.

domain's knowledge—a criterion now required for reference terminologies. Reference terminologies comprehensively and rigorously define the concepts and expressions within a biomedical domain, including interrelationships among concepts.[5,6] However, uniform, standard evaluation strategies relevant to reference terminologies may not translate well in the functional assessment of other types of terminologies.[1] The authors believe that intended usages must drive the evaluation of clinical terminologies to optimize the specificity and generality of findings.[21,22] In particular, evaluation of clinical interface terminologies should focus on how well they support efficient data entry and data review by intended users, such as healthcare providers.[2]

The authors previously identified salient desiderata for clinical interface terminologies,[2] including: a broad and richly-nuanced set of synonyms that accurately represent "natural language" phrases and expressions occurring in relevant biomedical discourse; a balance between pre-coordination and enabling post-coordination that maximizes clinician-users' efficiency in searching for concepts; the incorporation of assertional medical knowledge that links related concepts to one another, including to potential modifiers; a mapping to formal semantic structures; and the independence of interface terminologies from software applications that implement them. The authors describe below operational methods to assess the functional utility of a given interface terminology.

Model Description—Measurable Attributes for Interface Terminologies

Concept Coverage, Term Accuracy and Term Expressivity

As the initial step in evaluating a terminology, one should determine whether that terminology represents the knowledge-related entities in the domain it purports to cover.[6,17–19,23] For interface terminologies, this step should include evaluation of three measures: concept coverage, term accuracy, and term expressivity. The statistic "concept coverage" has been defined as the proportion of concepts or modifiers from a given domain that the terminology incorporates.[6,16–20] Operationally, investigators have assessed terminology coverage using as a "gold standard" expert-identified "important" words, terms, or phrases contained in natural language clinical documents spontaneously generated during care delivery. By "important," evaluators meant that the well-qualified, expert human reviewers believed that a given word, term, or phrase from the clinical document conveyed adequate meaning and that clinical necessity mandates that the item should be included in the terminology.[23–29] For example, a study by J. Campbell in 1997 reported that the then-current SNOMED International® terminology contained nearly 70% of the general medical concepts that investigators judged it should contain, based on a review of a sample set of clinical notes drawn from four large United States medical centers.[30]

Synonyms are words or phrases that provide interchangeable, alternative surface representations for words, phrases or the entirety of a formal concept. For example, the phrase "CAT scan" is a synonym (and abbreviation) for "computerized axial tomography image." Because synonyms can help users to find formal terms that match users' informal descriptions, the presence of adequate synonyms increases the usability (pragmatic utility) of an interface terminology.[31,32] While synonyms ideally have identical meanings, some terminologies define as synonyms two terms that have slightly different meanings.[33] The authors have previously proposed the metrics accuracy and expressivity to quantify the adequacy of a terminology system's synonyms.[2] Synonym *accuracy* reflects how well the meaning of a term's designated synonym corresponds with the meaning of the original term. For example, "anterior chest pain" is a more accurate synonym for the target concept, "substernal chest pain" than is "lateral chest pain." However, neither anterior nor lateral chest pain accurately matches the target concept, since both precordial and parasternal pain are also "anterior." Thus, synonym accuracy is determined by whether each component of a target term is represented accurately in a proposed synonymous term.

By contrast, term *expressivity* reflects how well a synonym's *semantic character* matches the words in the phrase it is meant to represent. Expressivity is more judgmental in nature. Semantic character consists of a qualified domain expert's subjective impression about the narrative flavor, the implicit clinical urgency, and the specificity of meaning conveyed by the words and the word order in a given clinical phrase.[2] For example, consider two arbitrary terminology systems—one that represents chest pain formally as being either "sharp" or "dull" and the other, which represents chest pain formally as being "sharp," "stabbing," or "dull." If a patient were to complain of "knife-like chest pain," an individual using the first hypothetical terminology might agree that the formal term "sharp" accurately matches the modifier "knife-like," but that it does not exactly match the patient's words. A care provider using the second hypothetical terminology would be more likely to consider, "stabbing pain" as equally accurate but more likely to be judged has having the same semantic character as "knife-like pain." Both hypothetical terminologies could be judged as covering the modifier, "knife-like" in the sense of accuracy (i.e., there is a canonical term that at least seems to represent the concept). However, the first terminology, containing only "sharp" but not "stabbing," is lacking in terms of expressivity compared to the latter terminology that contains "stabbing." Healthcare providers using an inaccurate or poorly expressive interface terminology would likely struggle and spend more time in entering their own "natural language" terms into a structured clinical application.

To measure the synonym accuracy and expressivity of a target terminology, project members would perform the following steps. First, they should obtain a reference set of clinical phrases that were previously and independently generated as clinical documentation during routine care delivery. In the past, to study terminology coverage, investigators have created such clinical phrase corpora from de-identified, pre-existing patient records.[24,30] Second, project members, studying, for example, a terminology for diagnostic findings, would employ appropriately qualified reviewers (e.g., experienced clinicians for the current example) to extract from the reference set of records any phrases containing "a statement by a clinician of a diagnostic impression or finding" (K. Campbell).[34] Third, project members would attempt to match terms from the target termi-

nology under study to the concepts extracted from the reference set of records. Fourth, for cases in which the target terminology contains multiple synonyms for a clinical concept, the project members should select what they judge to be the best-matching synonym. Fifth, project members should rate each of the selected "best-fit" synonyms, answering the questions, "does the selected synonym have the same general meaning as the clinical phrase?" for accuracy, and "does the target terminology's wording match the general understanding of the meaning of the synonym from the reference set?" for expressivity. Sixth, project members would calculate the overall percentage of clinical phrases that the target terminology accurately captured with good expressivity. Because this evaluation (as well as those that follow) involves subjective measurements, it is possible that inter-reviewer disagreements may occur. Investigators should employ multiple reviewers, then either calculate inter-reviewer agreement or address disagreement using consensus building methods, such as those employed in prior terminology evaluations.[31,35,36]

Principled Term Construction within a Terminology—Syntactic Consistency

Most terminologies present multiple acceptable ways of expressing a given concept. For example, the National Library of Medicine's Unified Medical Language System (UMLS)[37] terms have a preferred, or canonical (CUI) format, and multiple alternatives, represented by lexical unique identifier (LUI) and string unique identifier (SUI) strings linked to the given UMLS CUI concept. As has been the case for the UMLS, terminology maintainers may designate a single term as the "preferred" representation of the concept. Terminologies that apply a consistent syntax to preferred terms often prove easier for a human user to use than terminologies with "random" syntax or word orderings. For example, by applying a heuristic that says clinical finding names should have anatomical location descriptors that precede pathophysiological condition descriptors which are in turn followed by less important "adjectival" modifiers—the terms "abdomen pain colicky" and "abdomen tenderness rebound localized" may be easier to find from within long term lists. More liberal, but more difficult to use, finding nomenclature construction rules might allow terms like "substernal burning chest pain" and "knife-like chest pain anteriorly." Similarly, a terminology containing medical syndrome names might represent preferred terms with either generic descriptive names (e.g., "Acute febrile mucocutaneous lymph node syndrome") or with eponyms (e.g., "Kawasaki disease" for the same syndrome). SNOMED CT, for example, uses the descriptive name, "Acute febrile mucocutaneous lymph node syndrome" as the preferred term for the concept "Kawasaki disease," while also using the eponym, "West syndrome" as the preferred term for the concept "infantile spasms." Likewise, SNOMED CT represents "Babinski reflex" using the preferred term, "extensor plantar response finding," but uses the eponym "Hoffman's reflex" as the preferred term for "involuntary flexing of the end of the thumb and index finger elicited by tapping on the third finger nail."

Syntactic consistency within a terminology, also called "natural language consistency,"[2] involves application of internally standardized principles for preferred concept construction and wording. Better syntactic consistency helps interface terminology users to find and to select efficiently the best-matching terms for their ideas. For example, a user searching for a concept that has an obscure eponym would have to search for both the eponym and several generic forms of expressing a finding to determine if a terminology actually represented the concept. As noted, the task is much simpler if the terminology has consistent preferred term wording constructs.

Evaluators should examine the preferred terms in an interface terminology to assess syntactic consistency, using as a stimulus clinical phrases taken from the previously described reference set of clinical records. From the terms in the interface terminology that best match those in the reference set, the evaluators should identify each interface terminology preferred term, and categorize it as having one or more of the following features:

1. Whether or not modifiers are consistently and sensibly pre-coordinated with similar concepts (e.g., in SNOMED CT, the concept "structure of brachial artery" is not pre-coordinated with any modifiers for laterality, while the term "structure of left axillary artery" in the same terminology is pre-coordinated with a modifier for laterality).
2. Whether similar pattern or sequence of concepts in pre-coordinated terms (e.g., SNOMED CT contains both "Parotid swelling" [i.e., an anatomic concept, followed by abnormal finding] and "Mass of parotid gland" [i.e., an anatomic concept, preceeded by abnormal finding]).
3. Whether or not the "normal" or "non-diseased" status of a concept (e.g., whether an abnormal finding concept is absent or present) can be uniformly represented or expressed using the terminology, and whether the terminology consistently includes concepts that allow expression of the status (e.g., in SNOMED CT, "cranial bruit" is not pre-coordinated with the modifier "present", while the similar concept "femoral bruit present" is).
4. Whether the terminology includes extraneous and non-natural words in term names per se (e.g., the semantic type indicated by the word "structure" is included in some SNOMED CT terms for some anatomic locations, including "Left upper arm structure" and "Structure of left lower leg").
5. Whether or not the terminology uses variable or contradictory grammatical constructions within a single term name or in similar related terms (e.g., Medcin® contains both the concept "difficulty breathing better with sitting up" which includes the modifier "better," an adjective, and the concept, "difficulty breathing worsens with sitting up" which includes the modifier, "worsens," a verb).
6. Whether preferred term concepts are consistently represented using available eponyms or by more generic descriptive terms.

To assess syntactic consistency, evaluators should determine, across multiple examples, the rates at which each of the above-listed forms of expression occurred among the preferred terms in the interface terminology. For example, the project team might find that 10% of interface-preferred terms are pre-coordinated and include a modifier for laterality, while the other 90% of clinical phrases (where lateral-

ity is relevant) do not directly express it, but instead require a post-coordinated laterality modifier.

Compositional Balance in an Interface Terminology

The degree to which terminology developers enumerate (i.e., pre-coordinate) or spell out all possible complex concepts for users *a priori* may impact terminology usability.[2] A menu of all complex concepts expressible within the terminology may increase the chances that a user will find a desired term, but an exhaustive list might increase the size of the terminology to the point that users experience difficulty searching through it. For example, consider a pharmacy formulary for oral medications that lists as primary entries pairs of all brand name formulations and their available strengths. Rather than simply looking for "hydrochlorothiazide tablets" in a system that listed generic drug names as primary entries, with synonyms for brand names, the user would have to search through all combinations of several dozen brand names with three different dosage strengths each—potentially as many as 50–100 items to choose from.

For terminologies in general, an alternative approach to pre-coordination allows the user select (or build up) complex concepts from multiple-choice "pick lists" as needed. Such post-coordination may increase terminology flexibility, but might also increase users' difficulty in applying the terminology consistently, since there might exist multiple ways to build a given complex concept from primitive "atomic" concepts. Similarly, there might exist ways to construct nonsensical concepts through unconstrained selections from multiple axes.[1,38,39] Consider, for example, the difference between selecting the pre-coordinated term, "chest pain substernal crushing," versus post-coordinating "chest pain" from a list of clinical pain templates, then selecting "substernal" from a list of possible chest pain anatomical sites, and then selecting "crushing" from a list of possible "chest pain characters." Balancing pre-coordination during terminology design and construction with the ability for users to post-coordinate concepts as needed may optimize terminology flexibility, ease of use and overall coverage.[39] Optimizing this balance, which the authors have previously called "compositional balance,"[2] can facilitate users' concept selection tasks by minimizing the effort required to compose complex concepts from more atomic concepts, or to search through long lists of fully defined pre-coordinated concepts.

Evaluators of interface terminologies can apply J. Campell's method[39] to quantify compositional balance by first measuring concepts' terminological *degrees of freedom*, defined as a the number of atomic concepts present in each complex pre-coordinated concept. Campbell operationally defined atomic concepts as the most general concepts from a reference terminology that could be used to compose a more complex concept in a mapped interface terminology. For example, the interface term "Chest pain on exertion" in SNOMED CT has three degrees of freedom because it can be composed from the three atomic concepts, "exertion," "chest" and "pain." (The authors note that the causal relationship represented between the concepts exertion and chest pain is often represented using a semantic linkage rather than with a concept; semantic linkages are discussed below). Evaluators can calculate the degrees of freedom for each concept in an interface terminology by mapping them to a relatively granular reference terminology (such as mapping concepts in Medcin to SNOMED CT) and identifying how many reference terms are required to represent each interface term, irrespective of whether the reference concepts are from different terminological axes. Evaluators then calculate descriptive statistics such as mean and interquartile ranges for a terminology's degrees of freedom for single sub-axes or across the entire terminology. The authors have previously speculated that an optimal number of degrees of freedom for a terminology exists that maximizes terminology usability.[2] The authors believe that a terminology-wide average of 3–5 degrees of freedom could optimize usability by simultaneously minimizing the need for the user to search through longer lists of pre-coordinated concepts and reducing the effort required to compose complex concepts from large numbers of axes. The point at which this compositional balance is achieved may vary on the basis of the interface terminology's intended use and clinical domain.

Assertional Knowledge in an Interface Terminology

Terminologies ideally incorporate definitional knowledge (also called "terminological knowledge" and "contextual knowledge").[6,15,17–19,40] Such knowledge specifies, for individual concepts, the structural relationships from them to other concepts. For example, in SNOMED CT, the concept "chest pain" is formally defined by three relationships: 1) is-a "finding of region of thorax"; 2) is-a "pain of truncal structure"; and 3) has-finding-site "thoracic structure." In addition to definitional knowledge, interface terminologies should include assertional knowledge, which describes each concept's relationships to other concepts and modifiers, such whether concepts are present or absent under certain circumstances (e.g., the finding of an S4 left atrial gallop cannot occur in the presence of atrial fibrillation, a clinical disorder). Assertional knowledge also indicates whether a given term is relevant to specified patient populations (e.g., pregnancy status is not relevant for males, for women who have undergone hysterectomies, or for post-menopausal women).[2,40,41] Other forms of assertional knowledge may indicate which modifiers commonly describe a given concept (e.g., one may characterize chest pain by its severity and by its response to certain known exacerbating or relieving factors). Examples of assertional knowledge-based relationships in an interface terminology include whether concepts have a normal state defined, and lists of relevant modifiers and associated concepts. Such information, when incorporated into interface terminologies, improves their usability for encoding loosely structured natural language phrases.[41–44]

Measuring an interface terminology's assertional knowledge involves the review of linkages among concepts and linkages between concepts and modifiers. To do this review, evaluators examine clinical documents and reference materials such as textbooks to discover assertional knowledge-based relationships among an interface terminology's concepts and modifiers. Evaluators then determine whether the interface terminology represents such relationships. For example, upon reading the statement, "the patient complains of severe substernal chest pain radiating to the jaw," the reviewers would search an interface terminology for

linkages that connect the concept "chest pain" with the modifier "severe," the anatomic location "substernal area," the concept "radiates to jaw" and the status modifier "present." Reviewers' judgments as to whether such links are necessary and whether they are adequately implemented are subjective, and may vary among reviewers. Evaluators should use previously validated methods either to calculate inter-reviewer agreement or to address disagreement using consensus building methods.[31,35,45,46] Evaluators should also recognize that assertional knowledge may be implied in the precoordinations that exist in highly compositional terminologies such as Medcin, or explicitly contained in formally defined linkages among concepts and modifiers.

Formal Semantic Structure

Explicitly describing the relationships among concepts in a terminology (whether it is pre-coordinated or allows post-coordination) helps to provide a formal picture of a knowledge domain, and can improve the suitability of the terminology for automated data storage, management and analysis.[34,39,47] For example, the relationship that defines the concept "severe chest pain" as a more specific version of the concept "chest pain" can be described by the "is-a" and "has-severity" relationships (i.e., "severe chest pain" is-a [type of] "chest pain," has-severity "severe"). These relationships are examples of description logic properties. Description logics formally specify the relationships that may exist among terminology concepts and modifiers to support algorithmic data storage, inferencing, subsumption, classification, management and analysis. While interface terminologies may benefit from providing assertional knowledge-based linkages among concepts and modifiers to support data acquisition, such terminologies need not necessarily include formal description logic-defined relationships among concepts. Instead of embedding description logic linkages directly within interface terminologies, developers instead can map interface terminology elements to reference terminologies that have them.[3,4,34] Elkin expanded on prior work performed by Masarie, Miller et al. developing possible UMLS representation schemes;[42] this work had suggested that revealing that direct external representations of the semantics implied in pre-coordinated concepts can improve concept mapping accuracy.[48]

Terminology evaluators should determine whether interface terminologies internally contain formal description logics, or if they map adequately to reference terminologies having formally defined description logic linkages. By "adequate" mapping, the authors mean that it is pragmatically straightforward to use the linkages in the related reference terminology as an adjunct within any clinical program that itself employs the interface terminology. When an interface terminology includes its own description logic linkages, the situation is more straightforward than when mapping must occur. For the case where interface terminologies rely on mapped reference terminologies for semantic structure, evaluators should determine the accuracy of the mappings. To do so requires that evaluators review concepts in the interface terminology and the corresponding concepts in the reference terminology and judge whether they have the same meaning, and also determine if the relationships in the reference terminology accurately and appropriately hold for concepts in the interface terminology. For example, one might judge whether is it better to represent, in the reference terminology, "chest pain" has-location "substernal" (for the pre-coordinated interface term, "substernal chest pain"), or whether the relationship "chest pain substernal" is-a "chest pain" is preferred, less good, or complimentary (such that both representations are required). Evaluators can then calculate inter-reviewer agreement or address disagreement using consensus building methods. They can then derive statistics for the overall proportion of mappings that they believe to be accurate. For those interface terminologies that contain their own description logic linkages, investigators should evaluate those linkages directly. Methods for evaluating description logic linkages, including testing description logic attributes such as coverage, accuracy and ambiguity, have been previously described.[29,49–51]

Support for Human Readability

Interface terminologies often assist clinicians and other users to access, read and understand previously encoded data. To do this, interface terminologies typically convert an application's internally encoded data into more colloquial display phrases and terms, and may take advantage of grammatical tags to facilitate natural language generation. For example, Vanderbilt's Quill structured note capture application uses an internally stored augmented transition network (ATN) to generate clinician-readable text from discrete data.[52] The technique of using ATNs for such purposes is widespread, and was previously used, for example, in the mid-1970s by Edward H. Shortliffe in developing MYCIN's question answering system,[53] in the early 1980s by Perry Miller in the "Attending" anesthesia plan critiquing system[54] and in the 1990s by Poon and Johnson[55], and by Lehman.[56] The interface terminology supporting Quill includes, for every concept and modifier, attributes that specify the preferred term and its grammatical part of speech. For example, the term "chest pain" is tagged as a noun, and the modifiers "severe" and "substernal" as adjectives. When a Quill user selects the interface terms, "chest pain," "severe," "substernal" and "present," the application's ATN generates the sentence, "Severe substernal chest pain is present." In this way, a Quill user documenting discrete data using an interface terminology simultaneously can generate a naturalistic, human-readable note.

Evaluating interface terminologies designed to support human readability entails examination of two aspects of data presentation. First, evaluators should identify and characterize how the application uses the interface terminology to present data. For example, consider whether the application displays preferred terms or user-selected terms, how it sequences terms and modifiers, whether the system utilizes grammatical parts of speech to optimize presentation, whether it presents information in synoptic (i.e., bulleted) or narrative format, and whether it depends on a specific proprietary software to display its output correctly. Second, evaluators should test whether representative clinician-users rate the systems' display of discrete and encoded clinical data as easy to read and interpret. To do this, clinicians judge whether the terminology displayed correctly represents the underlying data. Specifically, reviewers would need to judge whether the output is accurate, is readable, is clear, and has a "natural" feel to it.

Application Independence

Electronic Health Record (EHR) systems generally incorporate interface terminologies in one of two ways. First, the terminology may be integrated directly into an EHR system application (e.g., the user interface programming directly contains and displays selectable terminology components through menus of drop down boxes, buttons, list boxes, etc). Such selectable items are interface terms by definition, but they do not constitute a standalone interface terminology. Second, the terminology and the user interface system may exist as separate application components joined together via a parallel implementation or a service oriented architecture, with each component having with its own distinct attributes. This level of independence allows both the user interface and the interface terminology to impact usability. For example, Medcin, when used as an interface terminology with proprietary documentation tools developed by the same software company that maintains it, includes components that enhance its usability, including a "qualifier table" linking common modifiers to related concepts.

Distinguishing the effects of interface terminology attributes from effects of an application's user interface attributes requires measuring similar outcomes for multiple different interface terminologies and for a variety of user interfaces. Previous studies used formal usability evaluation methods to test interface terminologies as implemented in single computer applications. Few evaluations have examined the interactions between an application's user interface and its underlying interface terminology.[57–59] To study interface terminology application independence, evaluators should first determine if an interface terminology was designed to be application independent, and then if possible, evaluate how embedding it within different computer applications affects its usability, according to the methods outlined below.

Evaluating Usability

After measuring the attributes outlined above, evaluators should determine how well an interface terminology performs when used to complete sample, representative real-world tasks for which designers built the interface terminology. Such evaluations should apply methods developed by usability science[60] to determine how much effort users must expend to complete tasks, how often users fail to complete tasks, and how satisfied users are with the experience. Related metrics include: completeness, which assesses the proportion of tasks that a user can perform successfully using a terminology; correctness, which determines how accurately completed tasks were performed when compared to a gold standard; efficiency, which measures the number of steps, the amount of time, and the perceived effort required to complete a given task using an interface terminology; and, satisfaction, which rates user impressions after interacting with an interface terminology to complete tasks.

Quantitative outcomes also include the number of times users correctly and completely accomplished tasks, and the time and number of actions required to complete a task. More detailed measurable actions include the numbers of term searches, steps required to browse either between hierarchy axes or to more or less specific concepts within a terminology axis, composing two concepts into a single concept or selecting between synonyms; these actions can be studied by recording keyboard events and mouse clicks.[51,57,58] Well-described methodologies for gathering qualitative data[61–65] include audio- or videotaping subject feedback, open-ended in-depth interviews with subjects, and detailed failure analyses. For example, subjects can "think out loud" as they perform tasks with an interface terminology. This provides information that the investigators can correlate with video and keystroke action logs. In cases when subjects fail to enter a test phrase, the evaluator can "debrief" the subject to understand why the subject had difficulty. Additionally, for each modeling task, evaluators can ask users to articulate their satisfaction with the correctness and the completeness of the selected or composed term, either through spoken feedback or via agreement scales for statements such as: "the selected concept correctly models the clinical phrase" and "the selected concept completely models the clinical phrase."

A Model for Evaluating Interface Terminologies

Interface terminology evaluations should include at least one of the following two components, depending on how designers envisioned use of the terminology. For interface terminologies designed to capture clinical documentation from healthcare providers while documenting clinical encounters, investigators should measure the following attributes as described above: 1) term coverage, accuracy and expressivity, 2) the proportion of preferred terms that reflect consistent natural human language, 3) determination of an appropriate balance between pre-coordination and post-coordination, 4) incorporated assertional medical knowledge, and 5) mapping to or inclusion of a formal semantic structure. For interface terminologies designed to display internal, application-encoded data to clinical users, investigators should measure: 1) term coverage, accuracy and expressivity, 2) how well the terminology supports human readability, and 3) whether the terminology can be used across multiple computer applications. After completing the foregoing steps, investigators should next evaluate interface terminology's usability for real-world tasks, such as encoding or reviewing data, measuring as outcomes task completeness, correctness, efficiency, and user satisfaction.

Upon completing the steps above, investigators will have gathered a detailed set of attribute measurements relative to an interface terminology's coverage for clinical phrases. Specifically, for each clinical phrase evaluated, the measurements would consist of yes/no indicators for concept coverage, term accuracy and term expressivity, a numeric indicator for each of the syntactic forms that terminology preferred terms take, the number of degrees of freedom contained in the terminology's representation of the clinical phrase, and yes/no indicators for the presence of assertional knowledge. The authors do not know of any empiric research that could currently guide weighting of these attributes to derive an overall "figure of merit" score that predicts interface terminology usability. Therefore, investigators should directly report both usability evaluations and the correlated interface terminology attributes metrics to help in developing an eventual weighting scheme. The clinicians' usability ratings will include, as previously noted, the time required to code clinical phrases using an interface

terminology, an indicator of whether the modeling was a success or failure, the numbers of actions required to encode the clinical phrase, and the user satisfaction on a numeric scale. Together, these data permit a multivariate analysis with and without interaction terms to determine the relative impact of the individual terminological attributes and the usability outcomes.

Demonstration Through Examples

To demonstrate how one might use the proposed model to evaluate interface terminologies, the authors identified two examples. In the first example the authors apply the evaluation model to SNOMED CT, detailing at length its performance against a component of a clinical note. In the second example, the authors apply the evaluation model to Medcin to demonstrate specific outcomes that can be exposed by the model. The examples were obtained after approval from the Vanderbilt Institutional Review Board. The authors randomly selected deidentified clinical statements from transcribed physician notes generated during primary care clinic encounters.

Example 1.

The example comes from the medical record of an 80 to 85 year old female with a history of hypertension and abdominal surgery. The note contained the following statement: "She has had about three months of intermittent left lower quadrant pain off and on. It lasts for one or a few hours. It is not associated with other symptoms such as constipation, diarrhea, nausea, vomiting, hematuria, melena, or hematochezia." While the authors elected to evaluate SNOMED CT as an interface terminology for this example, they note that SNOMED CT developers did not create it solely to serve as an interface terminology (i.e., it is equally designed to serve as a reference terminology). Nevertheless, the College of American Pathologists has endorsed SNOMED CT as, "the universal health care terminology that makes health care knowledge usable and accessible wherever and whenever it is needed".[66] For this example, the authors used SNOMED CT version 20060131, as represented in the Spring, 2006 UMLS distribution. Two authors working independently as reviewers (i.e., STR and PLE) with adjudication by a third (SHB) attempted to represent the concepts contained in the clinical statement with SNOMED CT. For the current example, adjudication took place by three-way conferencing among these three reviewers for cases of disagreement or ambiguity.

As a first step, the author-reviewers identified clinical concepts from the clinical record segment. These included the central concept "abdominal pain" localized to the left lower quadrant, represented in the term "left lower quadrant pain", and further modified by the terms, "three months," "intermittent," "off and on," lasting for "one or a few hours" and that it is "not associated with other symptoms." The concepts that the abdominal pain is asserted to be unassociated with include those represented in the phrase by the terms, "constipation," "diarrhea," "nausea," "hematuria," "melena," and "hematochezia." Next, these reviewers determined whether SNOMED CT contained concepts and modifiers to represent the extracted concepts. Table 1 lists the concepts and modifiers found to represent the clinical phrase, and demonstrates that SNOMED CT had complete coverage of the phrase's concepts. In two cases, SNOMED CT contained more than one concept that could reasonably represent those from the clinical statement: the word "hematuria" could be covered by either of the SNOMED CT concepts, "hematuria" or "blood in urine," and the term "hematochezia" by either "hematochezia" or "blood in stool." In each of these cases, SNOMED CT contained two separate concepts with unique preferred terms that represented closely related (or even duplicated) clinical entities.

Table 1 ■ Sample Mapping to SNOMED CT Concepts*

Clinical Phrase Component	SNOMED CT Concept
three months	Three (qualifier value) {79605009} Month (qualifier value) {258706009}
intermittent	left lower quadrant pain
left lower quadrant pain	Left lower quadrant pain (finding) {301716002}
off and on	Intermittent (qualifier value) {7087005}
one or a few hours†	One (qualifier value) {38112003} Few (qualifier value) {57176003} hour (qualifier value) {258702006}
not associated with other symptoms†	Associated procedure (attribute) {363589002} Symptom (finding) {19019007} Other (qualifier value) {74964007}
constipation	Constipation (disorder) {14760008}‡
diarrhea	Diarrhea (finding) {62315008}
nausea	Nausea (finding) {73879007}
vomiting	Vomiting (disorder) {15387003}‡
hematuria	Blood in urine (finding) {34436003}
melena	Melena (disorder) {2901004}‡
hematochezia	Blood in stool (disorder) {72256005}‡

*The concepts contained in the clinical phrase, "She has had about three months of intermittent left lower quadrant pain off and on. It lasts for one or a few hours. It is not associated with other symptoms such as constipation, diarrhea, nausea, vomiting, hematuria, melena, or hematochezia." SNOMED CT version 20060131, as represented in the Spring, 2006 UMLS distribution was used in this example. Linking semantics were not considered in this example

†Represening the clinical phrase required a compositional expression in SNOMED CT.

‡Best SNOMED CT concept for representing the clinical phrase is a diagnosis (i.e., "disorder") rather than a symptom (i.e., "finding") concept.

As a next step, the reviewers determined the accuracy, expressivity and semantic approach for the SNOMED CT concepts identified as matching the clinical phrase. As above, all but two of the terms from the clinical note mapped to single SNOMED CT concepts having preferred terms exactly matching the words in the clinical phrase. For the two terms that mapped to multiple (potentially duplicated) SNOMED CT concepts, the choice of which terminology concept that the reviewers select to represent the clinical phrase may impact how to judge its accuracy and expressivity. For example, both concepts "hematuria" and "blood in urine" can be defined as blood contained in voided urine, and can therefore be judged as being accurate representations of the clinical phrase. The reviewers each speculated that the patient described in the clinical note was commenting on the absence apparent red-tinged urine as a symptom (rather than on the absence of red blood cells by the test

urine microscopy) when reviewing her symptoms with her healthcare provider. By contrast, SNOMED CT specifically defined "hematuria" as a finding in which red blood cells are detected in the urine using an objective test such as microscopy, and "blood in urine" as a clinical observation of blood-colored urine. In this case, the SNOMED CT concept "blood in urine" more accurately represents the dictated term, but has a different semantic character; by contrast, the SNOMED CT concept hematuria has the same semantic character (in this case, the same words) as the dictated term hematuria, but may not accurately represent the concept that the clinician intended to document.

In the current example, the reviewers also observed that SNOMED CT contained variation in its semantic approach to constructing some of the preferred terms. For example, SNOMED CT used the scientific name "melena" as the preferred term for the concept describing stool colored black by digested blood, but used descriptive or colloquial terms for other concepts, including "blood in urine" rather than hematuria, "vomiting" rather than "emesis," and "blood in stool" rather than hematochezia.

As a next step, the reviewers evaluated whether the SNOMED CT concepts identified above represented the assertional knowledge contained in the dictated clinical phrase. The healthcare provider generating this statement considered and documented the presence of abdominal pain together with modifiers for location, timing and course, and a list of other potentially associated symptoms. The facts that abdominal pain can be located in the left lower quadrant, last for a few hours, be intermittent in nature, and be associated with constipation, diarrhea, nausea, vomiting, hematuria, melena, and hematochezia all represent assertional knowledge. To evaluate whether SNOMED CT contained assertional knowledge to represent these details, the reviewers examined the relationships modeled around the concept "Abdominal Pain" in SNOMED CT. None of the modifiers or associated concepts present in the clinical statement had corresponding SNOMED CT relationships linked to the concept "abdominal pain." This can be contrasted with the SNOMED CT concept "blood in urine," which had relationships specified for time course, episodicity, severity, and the presence of pain. Both concepts, "abdominal pain" and "blood in urine" had SNOMED CT relationships specified to lists of diagnoses that could cause the symptom.

Next, the reviewers quantified the degrees of freedom for selected SNOMED CT concepts and modifiers. The concept "Left Lower Quadrant Pain" in the clinical phrase above is a compositional expression that contains the more general SNOMED CT concept "Abdominal Pain" and modifiers for location, "Left" and "Structure of lower abdominal quadrant." "Abdominal Pain" itself represents a composition of the two SNOMED CT concepts, "General finding of abdomen" and "Pain finding at anatomical site." In sum, the concept "Left Lower Quadrant Pain" contains three general concepts and two modifiers. While the concept "Blood in Urine," by contrast, is coded in SNOMED CT as an atomic concept, it can also be decomposed to the general concepts, "Abnormal urinary product" and "Hemorrhage," and then the concept "Abnormal urinary product" to "Urine finding" and "Normality Finding." Of the eight concepts from SNOMED CT that covered those in the clinical statement, four could be considered compositional expressions: "left lower quadrant pain," "hematuria," "melena," and "hematochezia."

From this example evaluating the use of SNOMED CT as an interface terminology to represent the clinical statement above, the reviewers observed that it had adequate coverage, presented some challenges in terms of characterizing its accuracy, and its syntactic consistency and assertional knowledge did not completely cover those present in the clinical statement. While able accurately to cover all eight concepts and the modifiers contained in the extracted clinical record segment, two SNOMED CT terms were different than those represented in the segment. For these two, expressivity was reduced. In terms of syntactic consistency, SNOMED CT used as preferred terms both descriptive and colloquial terms (e.g., "blood in urine" and "vomiting") and medical jargon terms (e.g., "melena") rather than adhering to a single uniform approach for preferred terms. None of the assertional knowledge-based relationships judged to be present in the clinical statement were encoded in SNOMED CT, although SNOMED CT did include some assertional knowledge related to "abdominal pain." The two reviewers attempting to represent the clinical statement identified the same concepts from SNOMED CT, and both had difficulty determining the best ones to represent the terms "hematuria" and "hematochezia," but both ultimately selected the same ones (i.e., "blood in urine" and "blood in stool").

As a last step, the reviewers attempting to model this phrase provided qualitative feedback relative to SNOMED CT's usability as an interface terminology for the clinical statement provided in the current example. SNOMED CT's usability could not be directly compared against a gold standard documentation method, such as dictation with human transcription, because the clinical statement was identified retrospectively (i.e., the reviewers did not generate the dictation and could not assess the usability of dictation). As a result, its correctness and completeness relative to dictation could not be assessed. However, the reviewers' qualitative impressions if SNOMED CT's usability could still be recorded and characterized relative to the quantitative measurements, above. In this case, when representing the clinical statement using SNOMED CT, the reviewers were generally satisfied that the terminology was complete and efficient to use. The reviewers provided no comments or feedback for the words that mapped accurately and expressively to terminology concepts.

For the two concepts from the clinical statement that did not have clear term/concept matches in SNOMED CT, hematuria and hematochezia, the reviewers questioned the correctness of the mappings, wondering whether SNOMED CT contained concept duplication that could lead to ambiguity. One reviewer commented, "Haematuria is felt to be ambiguous in SNOMED CT and is actually marked as such. It is considered a non-current concept and it contains no children. This is because the blood component could be White Cells or Red Cells or perhaps buffy coat. . .Blood in the Urine is also ambiguous but not listed as such in SNOMED CT." The other reviewer commented, "A human reviewer searching SNOMED CT could choose to map the phrase 'the patient had hematuria' either to any of the 'hematuria' concepts, given the ambiguity. If they were to encounter one

with a preferred term 'hematuria' and another with a preferred term 'blood in urine' then one might expect a variable approach to modeling." The process of disambiguating the correct concept before selecting the correct SNOMED CT concept to represent hematuria took more steps and time than did the process of representing the other concepts. This had the effect of reducing efficiency, both when measured by reviewers' perceptions and by the amount of time and number of steps required for searching SNOMED CT for the best fit concept.

Example 2.

For this example, the authors examined how well the interface terminology Medcin can represent a phrase used in a US Department of Veterans Affairs (VA) general medical evaluation template for documenting compensation and pension examinations.[51] In this example, the authors focus on the attributes concept coverage, term accuracy and expressivity and syntactic consistency, as described above. Medcin does not explicitly incorporate assertional medical knowledge-based linkages among concepts (although assertional knowledge may be implied by the compositional concepts it contains) or a formal semantic structure, so these are not evaluated in this example. For the current example, the authors selected the phrase, "Inspection of the Spine and Back—Etiology of postural or gait abnormality," from the VA template physical examination section. This phrase provides clinicians with a place to delineate any physical exam findings from a patient's back or spine that might explain abnormalities in his posture or gait. In this case, two external physician reviewers evaluated mappings to Medcin, with adjudication by two authors (PLE and STR) for cases of disagreement. No single Medcin concept exists to represent the entire expression contained in the VA template phrase. However, the reviewers agreed that the VA template phrase was made up of the five individual concepts: "Inspection of the Spine," "[Inspection of the] Back," "Etiology," "Postural [Abnormality]," and "Gait Abnormality" (with the elements in brackets implied by the phrase). The best available Medcin concepts selected to represent these are presented in Table 2.

Table 2 ■ Sample mapping to potential Medcin concepts*

Clinical Phrase Component	Medcin Concept
Inspection of the Spine	Musculoskeletal Exam—Thoracic Spine [169813]
	Musculoskeletal Exam—Cervical Spine [7964]
	Musculoskeletal Exam—Lumbar / Lumbosacral Spine [169812]
	Musculoskeletal Exam—Thoracolumbar Spine [7985]
and Back†	Examination Of The Back [202199]
Etiology	no match
Postural‡	Posture [8204]
Gait Abnormality	Abnormality of Walk [733]

*The concepts contained in the clinical phrase, "Inspection of the Spine and Back—Etiology of postural or gait abnormality"
†[Inspection of the] is implied.
‡[Abnormality] is implied.

From among the concepts contained in the VA template phrase, one had no match (i.e., "Etiology"), the reviewers agreed that one matched only a more general concept (i.e., "Inspection of the Back") and one matched multiple concepts that were simultaneously more and less granular (i.e., "Inspection of the Spine," discussed below). The concepts selected to cover "Inspection of the Spine," in particular, presented the reviewers difficulty. While the Medcin concepts found for this concept from the VA template phrase were more granular, each focusing on a specific part of the spine (e.g., thoracic spine or cervical spine), they were also less granular because they represented the physical exam for these spinal locations in general, rather than specifying the method of spinal examination (i.e., spinal inspection). Therefore, from among the five concepts in the VA template phrase, one was not covered at all, one could be partially covered by related concepts, one was partially covered by a more general concept and two were covered by concepts having the same level of detail. In no cases were the Medcin terms identical to the words contained in the VA template phrase. As a result, the reviewers scoring Medcin's accuracy and expressivity concluded that Medcin could accurately represent two of the five concepts, but judged that it had poor expressivity. The lack of expressivity in this example is was felt to be the result of both the requirement to create a compositional expression to cover the VA template phrase, and the absence of synonyms that match the terms in the phrase. In addition, Medcin does not specify a method to combine concepts into a compositional expression.

In evaluating syntactic consistency, the reviewers noted that Medcin applies three different approaches for constructing the terms in this example. For the concepts selected to represent "Inspection of the Spine," Medcin's terms were constructed using a formalism containing first the patient assessment component (i.e., musculoskeletal exam) followed by a dash, then followed by the anatomic location being evaluated (e.g., cervical spine). In contrast, the concept selected to represent "[Inspection of the] Back" was constructed using a simple English statement (i.e., "examination of the back"). The concept for "Gait Abnormality" used a third formalism in which the concept status (i.e., abnormality) was placed first, followed by the status' subject (i.e., walk). No single formalism appeared to apply for the term construction among the Medcin concepts selected to represent the VA template phrase.

Discussion

The current manuscript outlines a model for evaluating terminologies designed to support the interaction between humans and structured clinical data, such as structured clinical documentation. The proposed model delineates relevant terminology attributes, as well as a process for evaluating their impact on an actual documentation task. The model emphasizes those attributes that promote correct and efficient structuring of clinical data by human users, specifically as they relate to terms' semantic character and relationships between related concepts and other concepts or modifiers. A view of this model is in Figure 1.

The evaluation parameters outlined in the current model involve trade-offs. For example, the most accurate concept in an interface terminology for representing a given clinical

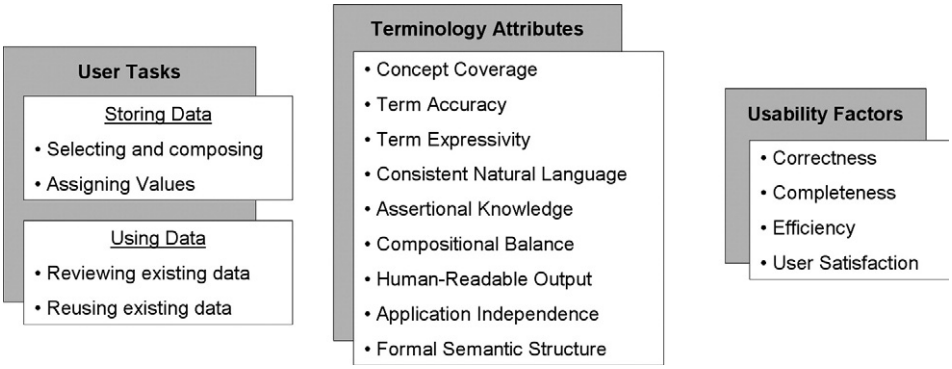

Figure 1. A Model for Evaluating Interface Terminologies. Users may use interface terminologies to accomplish two different tasks: encoding and storing clinical data and using existing data. The usability of interface terminologies can be measured using standard usability factors, including correctness, completeness, efficiency and satisfaction. These tasks and usability factors may be impacted by the terminology attributes.

phrase may not be the most expressive, and a terminology enforcing accuracy may reduce its usability. In one of the above examples, a note's author used the term "hematuria" generally to refer to the symptom of apparent blood in the urine while SNOMED CT used the term "blood in urine" to represent this entity, and had a different definition for the term "hematuria." To determine the best concept for representing a given clinical entity, the healthcare provider using the interface terminology would need to explore the differences between the two and determine which one carries both the meaning and the correct semantic character. In this case, the healthcare provider selecting a term must balance accuracy against efficiency and expressivity. Whether it is in the purview of interface terminologies to require highly accurate and precise encoding of clinical information is an open question. An inaccurate or imprecise interface terminology risks capturing incorrect data from the healthcare provider; an overly precise or accurate one risks being too rigid to be usable in general clinical practice.

When using a terminology to encode clinical information, the main tasks users perform include searching for and selecting the best concept to represent a clinical entity, setting its status (e.g., absent or present), and associating it with other related concepts or modifiers. Manually searching a terminology, either by keyword searches or by browsing through term lists, may be too inefficient for healthcare providers during clinical care delivery. Attributes improving users' abilities to find the best concept from within a terminology may enhance its usability. Many of the attributes proposed in the current evaluation model directly address the efficiency of term finding and selection. While powerful search tools in the user interface of an application using the interface terminology may help reduce the cognitive burden imposed by inadequate expressivity or language inconsistency, adequate synonymy and a predictable syntactic approach to forming terms can reduce the terminology's reliance on a given implementation environment.

It is possible that the proposed interface terminology attributes are incomplete or confounded by other factors that impact usability more than accuracy, expressivity and the identified assertional medical knowledge. It is also possible that investigators applying this model will demonstrate no association between the proposed attributes and the usability measures. Investigators can mitigate against the impact of this possibility by gathering qualitative feedback during the usability study. The qualitative data will allow investigators to identify reasons why they proposed attributes were not associated with usability. Possible reasons include a true absence of association, differential association based on previously unidentified sub-groups, and confounding by phrase, terminology or subject factors.

Conclusions

The authors have presented a model for evaluating interface terminologies in terms of how well they support the interaction between humans and structured concepts. A formal evaluation model will permit investigators to evaluate interface terminologies using a consistent and principled approach. The proposed model defines a series of measurable attributes that allow investigators to quantify how well the terminology can support efficient data entry and data review. Investigators can correlate these attributes with findings from usability evaluations of the interface terminology. Terminology developers and evaluators can apply this model to identify areas for improving interface terminologies.

References ■

1. Rector AL. Thesauri and formal classifications: terminologies for people and machines. Methods Inf Med. 1998;37(4–5):501–9.
2. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006;13(3):277–88.
3. Chute CG, Elkin PL, Sherertz DD, Tuttle MS. Desiderata for a clinical terminology server. Proc AMIA Symp. 1999:42–6.
4. Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. Proc AMIA Annu Fall Symp. 1997:640–4.
5. Campbell KE, Oliver DE, Spackman KA, Shortliffe EH. Representing thoughts, words, and things in the UMLS. J Am Med Inform Assoc. 1998;5(5):421–31.
6. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37(4–5):394–403.
7. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Toward a medical-concept representation language. The Canon Group. J Am Med Inform Assoc. 1994;1(3):207–17.
8. Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc. 2000;7(3):298–303.
9. Chute CG. The Copernican era of healthcare terminology: a re-centering of health information systems. Proc AMIA Symp. 1998:68–73.
10. Elkin PL, Brown SH, Carter J, et al. Guideline and quality indicators for development, purchase and use of controlled health vocabularies. Int J Med Inf. 2002;68(1–3):175–86.

11. Giuse NB, Giuse DA, Miller RA, et al. Evaluating consensus among physicians in medical knowledge base construction. Methods Inf Med. 1993;32(2):137–45.
12. Miller RA. Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc. 1994;1(1):8–27.
13. Miller RA, McNeil MA, Challinor SM, Masarie FE, Jr., Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project–status report. West J Med. 1986;145(6):816–22.
14. Rector AL, Nowlan WA, Glowinski A. Goals for concept representation in the GALEN project. Proc Annu Symp Comput Appl Med Care. 1993:414–8.
15. Rector AL, Nowlan WA, Kay S. Conceptual knowledge: the core of medical information systems. In: Lun KC, Deguolet P, Piemme TE, Rienhoff O, editors. Proceedings of the Seventh World Congress on Medical Informatics (MEDINFO '92); 1992; Geneva; 1992. p. 1420–6.
16. Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures. J Am Med Inform Assoc. 1998;5(6):503–10.
17. ISO/TS 17117:2002(E): Health Informatics—Controlled health terminology—Structure and high-level indicators: Technical Commitee ISO/TC 215, Health Informatics; 2002.
18. ISO 1087-1: Terminology work—Vocabulary Part 1: Theory and Application: Technical Commitee TC 37/SC 1; ISO Standards—Terminology (principles and coordination) 1996.
19. ISO 1087-2: Terminology work—Vocabulary Part 2: Computer applications: Technical Commitee TC 37/SC 3; ISO Standards—Computer applications for terminology; 1996.
20. ASTM 2087:2000: Standard Specification for Quality Indicators for Controlled Health Vocabularies: ASTM Committee E31 on Healthcare Informatics; 2002.
21. Winkelman WJ, Leonard KJ. Overcoming structural constraints to patient utilization of electronic medical records: a critical review and proposal for an evaluation framework. J Am Med Inform Assoc. 2004;11(2):151–61.
22. Friedman CP. Where's the science in medical informatics? J Am Med Inform Assoc. 1995;2(1):65–7.
23. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. J Am Med Inform Assoc. 1996;3(3):224–33.
24. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc. 1997;4(6):484–500.
25. Ruggieri AP, Elkin P, Chute CG. Representation by standard terminologies of health status concepts contained in two health status assessment instruments used in rheumatic disease management. Proc AMIA Symp. 2000:734–8.
26. Brown SH, Lincoln M, Hardenbrook S, et al. Derivation and evaluation of a document-naming nomenclature. J Am Med Inform Assoc. 2001;8(4):379–90.
27. Harris MR, Graves JR, Herrick LM, Elkin PL, Chute CG. The content coverage and organizational structure of terminologies: the example of postoperative pain. Proc AMIA Symp. 2000: 335–9.
28. Elkin PL, Bailey KR, Chute CG. A randomized controlled trial of automated term composition. Proc AMIA Symp. 1998:765–9.
29. Brown SH, Bauer BA, Wahner-Roedler DL, Elkin PL. Coverage of oncology drug indication concepts and compositional semantics by SNOMED-CT. AMIA Annu Symp Proc. 2003:115–9.
30. Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. J Am Med Inform Assoc. 1997;4(3):238–51.
31. Rosenbloom S, Awad J, Speroff T, et al. Adequacy of representation of the National Drug File Reference Terminology Physiologic Effects reference hierarchy for commonly prescribed medications. Proc AMIA. 2003, 569–573.
32. Cantor MN, Lussier YA. Putting Data Integration into Practice: Using Biomedical Terminologies to Add Structure to Existing Data Sources; Proc AMIA. 2003, 125–9.
33. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: An Exploration of Different Views of Synonymy and Quality of Editing. J Am Med Inform Assoc. 2005;2(4):486–94. (Epub 2005 Mar 31.)
34. Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Inform Assoc. 1994; 1(3):218–32.
35. Eagon JC, Hurdle JF, Lincoln MJ. Inter-rater reliability and review of the VA unresolved narratives. Proc AMIA Annu Fall Symp. 1996:130–4.
36. Elkin PL, Bailey KR, Ogren PV, Bauer BA, Chute CG. A randomized double-blind controlled trial of automated term dissection. Proc AMIA Symp. 1999:62–6.
37. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998;5(1):1–11.
38. Rassinoux AM, Miller RA, Baud RH, Scherrer JR. Compositional and enumerative designs for medical language representation. Proc AMIA Annu Fall Symp. 1997:620–4.
39. Campbell JR. Semantic features of an enterprise interface terminology for SNOMED RT. Medinfo. 2001;10(Pt 1):82–5.
40. Smart JF, Roux M. A model for medical knowledge representation application to the analysis of descriptive pathology reports. Methods Inf Med. 1995;34(4):352–60.
41. Chute CG, Elkin PL. A clinically derived terminology: qualification to reduction. Proc AMIA Annu Fall Symp. 1997:570–4.
42. Masarie FE, Jr., Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 1991;24(4):379–400.
43. Medical Subject Headings browser. Available at: http://www.nlm.nih.gov/mesh/MBrowser.html. Accessed Nov 4, 2003.
44. Wang AY, Barrett JW, Bentley T, et al. Mapping between SNOMED RT and Clinical terms version 3: a key component of the SNOMED CT development process. Proc AMIA Symp. 2001:741–5.
45. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002;35(2):99–110.
46. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.
47. Spackman KA. Normal forms for description logic expressions of clinical concepts in SNOMED RT. Proc AMIA Symp. 2001: 627–31.
48. Elkin PL, Brown SH, Lincoln MJ, Hogarth M, Rector A. A formal representation for messages containing compositional expressions. Int J Med Inf. 2003;71(2–3):89–102.
49. Sowa J. Knowledge Representation: Logical, Philosophical, and Computational Foundations: Logical, Philosophical, and Computational Foundations. Pacific Grove, CA: Brooks Cole Publishing Co.; 1999.
50. Bakken S, Warren J, Lundberg C, et al. An evaluation of the utility of the CEN categorical structure for nursing diagnoses as a terminology model for integrating nursing diagnosis concepts into SNOMED. Medinfo. 2001;10(Pt 1):151–5.
51. Brown SH, Elkin P, BA B, et al. SNOMED CT: Utility for a General Medical Evaluation Template. Proc AMIA Annu Fall Symp 2006.

52. Shultz EK, Rosenbloom ST, Kiepek WT, et al. Theater Style Demonstration - Quill: A Novel Approach to Structured Reporting. Proc AMIA Annu Fall Symp. 2003.
53. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res. 1975;8(4):303–20.
54. Miller PL. Critiquing anesthetic management: the "ATTENDING" computer system. Anesthesiology. 1983;58(4):362–9.
55. Poon AD, Johnson KB, Fagan LM. Augmented transition networks as a representation for knowledge-based history-taking systems. Proc Annu Symp Comput Appl Med Care. 1992:762–6.
56. Lehmann CU, Nguyen B, Kim GR, Johnson KB, Lehmann HP. Restricted natural language processing for case simulation tools. Proc AMIA Symp. 1999:575–9.
57. Poon AD, Fagan LM, Shortliffe EH. The PEN-Ivory project: exploring user-interface design for the selection of items from large controlled vocabularies of medicine. J Am Med Inform Assoc. 1996;3(2):168–83.
58. McKnight LK, Elkin PL, Ogren PV, Chute CG. Barriers to the clinical implementation of compositionality. Proc AMIA Symp. 1999:320–4.
59. Cimino JJ, Patel VL, Kushniruk AW. Studying the human-computer-terminology interface. J Am Med Inform Assoc. 2001; 8(2):163–73.
60. Neilson J. Usability Engineering. Boston, MA: Academic Press; 1993.
61. Ash JS, Berg M, Coiera E. Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Informations System Related Errors. JAMIA Pre-Print 2003.
62. Kvale S. Interviews: an introduction to qualitative research interviewing. Thousand Oaks, Calif.: Sage Publications; 1996.
63. Lofland J, Lofland J. Analyzing social settings: a guide to qualitative observation and analysis. 4th ed. Belmont, CA: Wadsworth/Thomson Learning; 2006.
64. May KA. Interviewing techniques in qualitative research: Concerns and challenges. In: Morse JM (editor) Qualitative nursing research: a contemporary dialogue (revised edition) Newbury Park, CA: Sage Publications; 1991:188–210.
65. Patton MQ. Qualitative research and evaluation methods (3rd edition). Thousand Oaks, Calif.: Sage Publications; 2002.
66. SNOMED CT. Available at: www.snomed.com. Accessed April 17, 2007.