



Standard Specification for Quality Indicators for Controlled Health Vocabularies¹

This standard is issued under the fixed designation E 2087; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

INTRODUCTION

In 1839, William Farr stated in his First Annual Report of the Registrar-General of Births, Deaths, and Marriages in England, “The nomenclature is of as much importance in this department of inquiry, as weights and measures in the physical sciences, and should be settled without delay.” Since that time this theme has been heard resounding from an increasingly large group of scientists (see [Appendix X1](#)). Today, the need for controlled vocabularies to support health record systems has been widely recognized (see Specification [E 1238](#), Guide [E 1239](#), Guide [E 1384](#), Specification [E 1633](#), and [EN 12017](#)). Controlled vocabularies provide systems with the means to aggregate data. This aggregation of data can be done at multiple levels of granularity and therefore can enhance the clinical retrieval of a problem oriented record, data pertaining to a classification for billing purposes, or outcomes data for a given population. Maintenance of large-scale vocabularies has become a burdensome problem as the size of term sets has escalated ([IS 15188](#)). Without a well-structured backbone, large-scale vocabularies cannot scale to provide the level of interoperability required by today’s complex electronic health record applications.

The solution rests with standards (1).² Over the past ten or more years, Medical Informatics researchers have been studying controlled vocabulary issues directly. They have examined the structure and content of existing vocabularies to determine why they seem unsuitable for particular needs, and they have proposed solutions. In some cases, proposed solutions have been carried forward into practice and new experience has been gained (2). As we prepare to enter the twenty-first century, it seems appropriate to pause to reflect on this experience, and publish a standard set of goals for the development of comparable, reusable, multipurpose, and maintainable controlled health vocabularies ([IS 12200](#), [IS 12620](#)) (3).

This specification represents the initial input taken from the ANSI-HISB Framework Paper by Chute, et al (4), the Desiderata from Cimino (3), the ToMeLo Architecture and Terminology Paper by Rossi-Mori and Zanstra, and the Compositionality Paper by Elkin, et al (5). Other useful references include, “GALEN Generalized Architecture for Language, Encyclopedias and Nomenclatures in Medicine: Univ. of Manchester” (6, 7) and “Unified Medical Language System (UMLS) Knowledge Sources” (8).

1. Scope

1.1 This specification covers the documentation of the principal notions necessary and sufficient to assign value to a controlled health vocabulary. This specification will serve as a guide for governments, funding agencies, terminology developers, terminology integration organizations, and the purchasers and users of controlled health terminology systems working toward improved terminological development and recognition

of value in a controlled health vocabulary. It is applicable to all areas of health care about which information is kept or utilized. It is intended to complement and utilize those notions already identified by other national and international standards bodies.

1.2 This specification will provide vocabulary developers and authors with the guidelines needed to construct useful, maintainable controlled health vocabularies. These tenets do not attempt to specify all of the richness that can be incorporated into a health terminology. However this specification does specify the minimal requirements, which, if not adhered to, will ensure that the vocabulary will have limited generalizability and will be very difficult, if not impossible, to maintain. This specification will provide terminology developers with a sturdy starting point for the development of controlled health vocabularies. This foundation serves as the

¹ This specification is under the jurisdiction of ASTM Committee E31 on Healthcare Informatics and is the direct responsibility of Subcommittee E31.35 on Healthcare Data Analysis.

Current edition approved May 10, 2000. Published July 2000.

² The boldface numbers in parentheses refer to the list of references at the end of this standard.

basis from which vocabulary developers will build robust, large-scale, reliable and maintainable terminologies.

1.3 This specification explicitly does not refer to classifications or coding systems (for example, a simple list of pairs of rubrics and codes) that are not designed to be used clinically.

2. Referenced Documents

2.1 ASTM Standards:

- E 1238 Specification for Transferring Clinical Observations Between Independent Computer Systems³
- E 1239 Guide for Description of Reservation/Registration-Admission, Discharge, Transfer (R-ADT) Systems for Automated Patient Care Information Systems³
- E 1284 Guide for Construction of a Clinical Nomenclature for Support of Electronic Health Records³
- E 1384 Guide for Content and Structure of the Electronic Health Record (EHR)³
- E 1633 Specification for Coded Values Used in the Electronic Health Record³
- E 1712 Specification for Representing Clinical Laboratory Procedure and Analyte Names³

2.2 Other Standards:

- ISO/DIS 860 International Harmonization of Concepts and Terms
- EN 12017 Medical Informatics—Vocabulary
- EN 12264 Medical Informatics—Categorical Structure of Syntax of Concepts—Model for Representation of Semantics
- ICD-9-CM
- IS 704 Principles and Methods of Terminology
- IS 1087-1 Terminology—Vocabulary—Part 1: Theory and Application
- IS 1087-2 Terminology—Vocabulary—Part 2: Computer Applications
- IS 11179-3 Terminology—Data Registries
- IS 12200 Terminology—Computer Applications—Machine Readable Terminology Interchange Format
- IS 12620 Terminology—Computer Applications—Data Categories
- IS 15188 Project Management for Terminology Standardization
- IS 2382-4 Information Technology—Vocabulary—Part 4: Organization of Data
- ISO TR 9789 Guidelines for the Organization and Representation of Data Elements for Data Interchange—Coding Methods and Principles

3. General Information

3.1 Basic characteristics of a terminology influence its utility and appropriateness in clinical applications.

3.1.1 *Concept Orientation* (3)—The basic unit of a vocabulary must be a concept, which is the embodiment of some specific meaning and not a code or a character string. Representations of a concept must correspond to one and only one meaning, and in a well-ordered vocabulary only one concept may have that same meaning (ISO/DIS 860). However, mul-

iple terms (linguistic representations) may have the same meaning if they are explicit representations of the same concept. This implies non-redundancy, non-ambiguity, and non-vagueness.

3.1.1.1 *Non-redundancy*—Terminologies must be internally consistent. There must not be more than one concept in the terminology with the same meaning (IS 704, Guide E 1284). This does not exclude synonymy; rather, it requires that this be explicitly represented.

3.1.1.2 *Non-Ambiguity*—No concept should have two or more meanings. However an entry term (some authors have referred to this as an “interface terminology”) can point to more than one concept (for example, MI as a myocardial infarction and mitral insufficiency).

3.1.1.3 *Non-Vagueness*—Concept names must be context free (some authors have referred to this as “context laden”). For example “diabetes mellitus” should not have the child concept “well controlled,” instead the child concept’s name should be “diabetes mellitus, well controlled.”

3.2 *Purpose and Scope*—Any controlled vocabulary must have its purpose and scope clearly stated in operational terms so that its fitness for particular purposes can be assessed and evaluated (IS 15188). Where appropriate, it may be useful to illustrate the scope by examples or “use cases” as in database models and other specification tools. Criteria such as coverage and comprehensiveness can only be judged relative to the intended use and scope. For example, a vocabulary might be comprehensive and detailed enough for general practice with respect to cardiovascular signs, symptoms, and disorders, but inadequate to a specialist cardiology or cardiothoracic surgery unit. Conversely, a vocabulary sufficiently detailed to cope with cardiology and cardiothoracic surgery might be totally impractical in general practice.

3.3 *Coverage* (3)—Each segment of the healthcare process must have explicit in-depth coverage and not rely on broad summary categories that lump specific clinical concepts together. For example, it is often important to distinguish specific diagnosis from categories presently labeled Not Elsewhere Classified (NEC), or to differentiate disease severity such as indolent prostate cancer from widely metastatic disease. The extent to which the depth of coverage is incomplete must be explicitly specified for each domain (scope) and purpose as indicated in 3.2.

3.4 *Comprehensiveness* (9)—All segments of the healthcare process, such as physical findings, risk factors, or functional status, must be addressed for all related disciplines, across the breadth of medicine, surgery, nursing and dentistry. This criterion applies because decision support, risk adjustment, outcomes research, and useful guidelines require more than diagnoses and procedures. Examples include existing AHCPR guidelines and the HCFA mortality model. The extent to which the degree of comprehensiveness is incomplete must be explicitly specified for each domain (scope) and purpose as indicated in 3.2.

3.5 *Mapping* (10)—Government and payers mandate the form and classification schema for much clinical data exchange. Thus, comprehensive and detailed representations of patient data within computer-based patient record systems

³ Annual Book of ASTM Standards, Vol 14.01.

able to be mapped to those classifications, such as **ICD-9-CM**. This need for multiple granularities is needed for clinical health care as well (**ISO TR 9789**). For example, an endocrinologist may specify more detail about a patient's diabetes mellitus than a generalist working in an urgent care setting, even though both may be caring for the same patient. The degree to which the terminology is isolated from other classifications must be explicitly stated.

3.6 Systematic Definitions (4)—In order for users of the vocabulary to be certain that the meaning that they assign to concepts is identical to the meaning which the authors of the vocabulary have assigned, these definitions will need to be explicit and available to the users. Further, as relationships are built into vocabularies, multiple authors will need these definitions to ensure consistency in authorship.

3.7 Formal Definitions—A compositional system should contain formal definitions for non-atomic concepts and formal rules for inferring subsumption from the definitions (Specification **E 1712**).

3.8 Explicitness of Relations—The logical definition of subsumption should be defined. The formal behavior of all links/relations/attributes should be explicitly defined. The primary hierarchical relation should be subsumption (“kind of”) as defined by logical implication: “B is a kind of A” means “All Bs are As.” If a looser meaning such as “broader than/narrower than” is used, it should be explicitly stated.

3.9 Reference Terminology—The set of canonical concepts, their structure, relationships, and, if present, their systematic and formal definitions. These features define the core of the controlled health terminology.

3.10 Atomic Reference Terminology—A reference terminology consisting of only atomic concepts and their systematic and formal definitions. In this type of reference terminology, no two or more concepts can be combined to create a composite expression as the same meaning as any other single concept contained in the atomic reference terminology.

3.11 Colloquial Terminology—The set of terms that consist of commonly used entry points and which map to one or more canonical terms within the vocabulary. These have been called “entry terms” or “interface terminologies” by different authors.

4. Structure of the Terminology Model

4.1 Terminology structures determine the ease with which practical and useful interfaces for term navigation, entry, or retrieval can be supported (**IS 704**, **IS 1087-1**, **EN 12264**). Terminologies that do not currently meet these criteria can be in compliance with this specification by putting mechanisms in place to move toward these goals.

4.2 Compositionality—Atomic concepts must be able to be combined to create composite concepts (**11**). A concept is a notion represented by language, which identifies one idea. For example, “colon cancer” comprises “neoplasm, malignant” and “large bowel” as atomic components. In a compositional system, concept representations can be divided into atomic and composite concept representations. Composite concept representations can be further divided into “named pre-coordinated concept representations” and “post coordinated representation expressions.” Within a composite concept, it may be possible to separate the constituents into three categories: the kernel

concept, modifier concept, and qualifier (also called “status”) concept. These terms are being specifically defined in a document on meta-terminology currently being written under the auspices of ISO TC 215 Working Group 3.

NOTE 1—The term “concept” in this specification is used to refer to the representation of a concept rather than the thought itself.

4.2.1 Atomic Concept—A representation of a concept that is not composed of other simpler concept representations within a particular terminology. In many cases “atomic concepts” will correspond to what philosophers call “natural kinds.” Such entities cannot be meaningfully decomposed. Concepts should be separable into their constituent components, to the extent that it is practical. These concepts should form the root basis of all concepts. For example, in the UMLS Metathesaurus, colon is a synonym for large bowel, and cancer is a synonym for neoplasm, malignant. Colon cancer is non-atomic, since it can be broken down into “large bowel” and “neoplasm, malignant.” Each of these two more atomic terms has a separate and unique Concept Unique Identifier (CUI).

4.2.2 Composite Concept—A concept composed as an expression made up of atomic concepts linked by semantic representations (such as roles, attributes, or links).

4.2.2.1 Pre-coordinated Concept—An entity that can be broken into parts without loss of meaning (can be meaningfully decomposed) when the atomic concepts are examined in aggregate. These are representations, which are considered single concepts within the host vocabulary. Ideally, these concepts should have their equivalent composite concepts explicitly defined within the vocabulary (that is, the vocabulary should be normalized for content). For example, colon cancer is non-atomic, however it has a single CUI, which means to the Metathesaurus that it represents a single concept. It has the same status in the vocabulary as the site “large bowel” and the diagnosis “neoplasm, malignant.”

4.2.2.2 Post-coordinated Concept—A composite concept is not pre-coordinated and therefore must be represented as an expression of multiple concepts using the representation language. This is the attempt of a system to construct a set of concepts from within a controlled vocabulary to more completely represent a user's query. For example, the concept “bacterial effusion, left knee” is not a unique term within the SNOMED-RT terminology. It represents a clinical concept that some patient has an infected left knee joint. As it cannot be represented by a single concept identifier, to fully capture the intended meaning a system would need to build a representation from multiple concept identifiers or lose information to free text.

4.2.3 Types of Atomic and Pre-coordinated Concepts—We can classify unique concept representations within a vocabulary into at least three distinct types: kernel concepts, modifiers, and qualifiers (which contain status concepts). This separation allows user interfaces to provide more readable and therefore more useful presentations of composite concepts.

4.2.3.1 Kernel Concept—An atomic or pre-coordinated concept, which represents one of the one or more main concepts within a pre-coordinated or post-coordinated composition.