Clinical Terminology Support for a National Ambulatory Practice Outcomes Research Network

Thomas N. Ricciardi[a,b], Michael I. Lieberman[a,b], Michael G. Kahn[a], F.E. "Chip" Masarie, Jr.[b,c]
[a] GE Healthcare Technologies, Waukesha Wisconsin, USA
[b] Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland Oregon USA
[c] Masarie Consulting, Portland Oregon USA

Abstract

The Medical Quality Improvement Consortium (MQIC) is a nationwide collaboration of 74 healthcare delivery systems, consisting of 3755 clinicians, who contribute de-identified clinical data from the same commercial electronic medical record (EMR) for quality reporting, outcomes research and clinical research in public health and practice benchmarking. Despite the existence of a common, centrally-managed, shared terminology for core concepts (medications, problem lists, observation names), a substantial "back-end" information management process is required to ensure terminology and data harmonization for creating multi-facility clinically-acceptable queries and comparable results. We describe the information architecture created to support terminology harmonization across this data-sharing consortium and discuss the implications for large scale data sharing envisioned by proponents for the national adoption of ambulatory EMR systems.

Keywords:
Data Warehouse, Terminology Harmonization, Ambulatory Electronic Medical Records

Introduction

Proponents for the national adoption of electronic medical records (EMRs) cite numerous benefits of widespread EMR adoption both to direct patient care and to the healthcare delivery system [1-3]. Implicit in these projected benefits is the ability for users in different practice settings using different EMR systems to share data so that clinically meaningful clinical states and outcomes can be combined into comparable numerators and denominators. Without the ability to combine EMR data from different providers, a key national benefit for the substantial financial investment in EMRs will not be realized.

The Medical Quality Improvement Consortium (MQIC) consists of 79 users of a commercial ambulatory medical record system (Centricity Physician Office, GE Healthcare Information Technologies, Waukesha, WI) who have agreed to contribute de-identified detailed clinical data into a national data warehouse to support multi-institutional clinical research and practice benchmarking. Figure 1 illustrates the general architecture for the national ambulatory data warehouse. Processes that execute nightly within the local practice are responsible for extracting clinical data from the local EMR database, for de-identifying clinical data, and for transferring the local data extract to the national data center. Processes that execute within the national data center are responsible for data cleaning, aggregation, query processing, and results presentation.

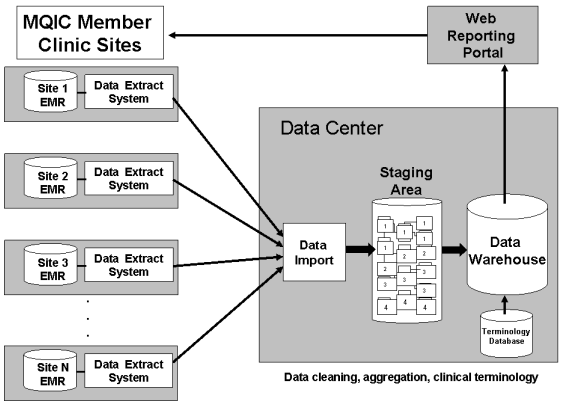


Figure 1: MQIC data warehouse infrastructure.

Although the processes that extract, de-identify and transfer data from the local ambulatory practice are important components of the MQIC information management architecture, we focus here on the terminology harmonization and data scrubbing processes that execute in the national data warehouse. These latter components provide the most insight into the effort required to create a broad-based national ambulatory practice research network.

Clinical Terminology Management

To aggregate data from across enterprises, a system was put in place to control clinical terminology in the data warehouse and to link user-entered custom terms to reference external terminologies (Figure 2: Clinical terminology subsystem.).

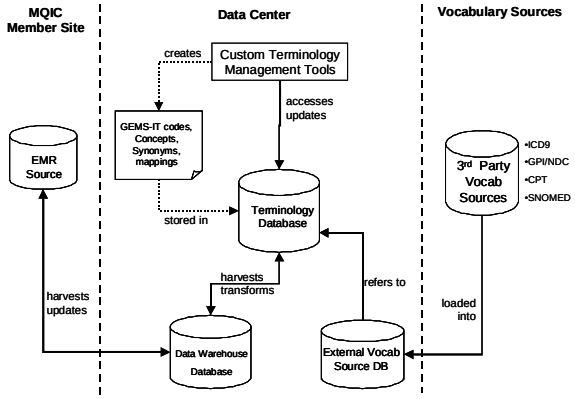


Figure 2: Clinical terminology subsystem.

EMR users document chief complaints, problems, medications, and orders by selecting a previously defined controlled term, by modifying a previously defined controlled term or by entering an uncontrolled free-text string. We call the controlled terminology developed for these clinical concepts a *user-interface terminology*. The same user-interface terminology is used by all local instances of the EMR and is managed from a single master terminology database maintained by the commercial vendor. The user-interface terminology has been developed heuristically by analyzing the free-text strings entered by clinicians to document problems, medications, orders, and chief complaints in the EMR.

The user-interface terminology currently contains over 650,000 strings mapped to over 86,000 concepts. Concepts are also mapped to administrative classifications such as ICD-9-CM and CPT-4. In addition, the user-interface terminology has been partially mapped to a reference terminology (SNOMED-CT), which allows aggregation of concepts for reporting purposes [4].

A second vocabulary that also is controlled centrally manages the attributes of clinical encounters that can be recorded as discrete elements, such as laboratory test names and vital signs (generically called "observations"). This second set of controlled terms, called the *observations terminology*, is used as the containers for observation values. Unlike the user-interface terminology, the observations terminology cannot be extended dynamically by the end-user using free-text strings. Observation terms are created by the central terminology service based on user requests for new EMR observations that are displayed on multiple flow sheets or are items that clinicians or managers want to use in practice-performance or clinical-quality reports. The current observations terminology consists of over 10,000 observation codes.

The final aspect of the EMR relevant to the national data warehouse is the data values that are entered for observations. Although user sites can control what values are used to instantiate observation terms through data entry forms, the EMR application is permissive in how users can enter values for observations to ensure that users have the flexibility to say what is required in a manner that is familiar and comfortable for each practitioner.

Thus, the EMR system's use of controlled terminology varies from requiring only controlled terms (observations), to a mixed model of controlled terms and free-text (user-interface), to an unconstrained terminology (observation values).

Clinical Terminology Harmonization

Table 1 provides examples of user-entered strings that should and should not be mapped to a previously defined controlled term in the user-interface master terminology.

Table 1: User terms – Matches & Non-matches

User-Term	Concept	Reason
Appropriate to Match		
Patient short of breath	Dyspnea	Synonym
Oosteeoarthiritis	Osteoarthritis	Misspelling
Degen joint disease	Osteoarthritis	Synonyms
COPD	Chronic obstructive pulmonary disease	Acronym
Inappropriate to Match		
Arthritis	Osteoarthritis	Requires inference
Osteoarthritis, knee	Osteoarthritis	Needs more granular concept
Periph neuropathy, H/O	Peripheral neuropathy	"history of" terms semantically different

Raw terms from all participating EMR locations are imported on a nightly cycle into a work queue, which becomes the working set of unprocessed terms. A user-entered term from the work queue that is already in the vocabulary database will be automatically mapped to the correct concept. Straightforward problem list strings, correctly spelled and free of superfluous text, always fall into this category. Terms with common misspellings are found by manually querying the work queue using text fragments or other substrings deemed effective by a vocabulary specialist. The most unusual strings have the greatest likelihood of being unmapped. When a user-entered term does not match any existing concepts and clinical knowledge (including research) determines the need for a discrete concept to preserve the intended meaning of the string, a new concept is created. In order to preserve as much clinical detail as was present in the original user-entered string, terms are not grouped into higher-level concepts ("right knee pain" is not mapped to "knee pain"). If the concept "knee pain, right" did not exist in the vocabulary database, it would be created, using consistent naming conventions, rules regarding expression of laterality, severity, etc., and one or more alternative names would be included in the new definition. The new concept would then be assigned the appropriate ICD-9, CPT, or medication code.

For the purpose of capturing even general types of information, a number of very broad concepts, such as "screening, breast cancer", "pain", "behavior problems", have been created. A small percentage of strings lack sufficient information to identify as concepts or are too vague to map. These user-entered strings are sent to an un-mappable collection and are not assigned an existing code nor is a new term created. Examples of un-mappable terms include "lab panel", "knee problem", and "on medication."

In addition to mapping concepts, data values also require harmonization prior to entering into the data warehouse. Table 2 illustrates actual values entered for systolic blood pressures and the results of the data scrubbing process. Entered values that are clearly numeric can easily translate to meaningful cleaned values; entered values with appended text are also easily cleaned. Entered values that are unprocessable, for example blood pressure measurements such as "refused" or "not done", receive a value of -1. Entered values that fall below or above a broadly defined range of plausible values receive a cleaned value of -2 and -3 respectively.

Interpreting textual data values also requires harmonization. Because MQIC is a collaboration of many different institutions, specific data fields can have a wide variety of values. For example, the data field titled *Diabetic Eye Exam* may be used to record the date of the exam in one institution, the impression from a consult at another institution, and one of a predefined set of values at a third institution. The master terminology contains a set of acceptable values for each data element, and then maps what the EMR users have entered to one of this controlled set.

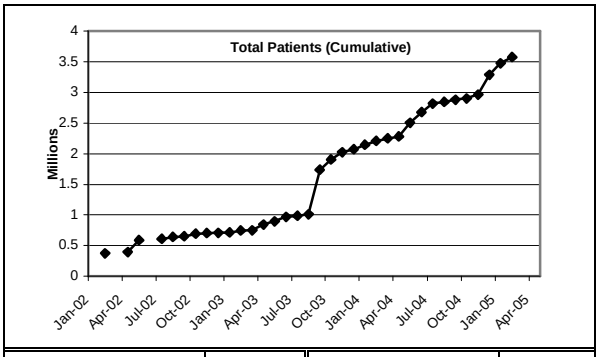
Table 2: Data scrubbing results for systolic blood pressure values.

Entered Value	Cleaned Value
120	120
110 RT	110
130 L LGE CUFF	130
Refused	-1
3	-2
350	-3

Results

Table 3 presents the distribution of practices participating in the MQIC data warehouse as of March 2005. As shown in Table 3, the MQIC warehouse is a reasonable first approximation for a national ambulatory practice data warehouse.

Table 3: Features of the MQIC network.



Practice Size	N	Practice Size	N
1 to 4 MDs	24	25 to 99 MDs	17
5 to 24 MDs	25	100+ MDs	8
Generalists	2112	**Specialists**	1571

Figure 3 plots key monthly statistics for mapping user-entered problem list strings and Figure 4 plots user-entered observation values. Both plots provide the absolute number of strings processed (lower diamonds) and the percent of strings successfully matched without manual intervention (upper squares). Figure 3 shows that for user-entered problem list strings, the matching frequency continues to rise slowly to nearly 90%. Large spikes in the number of new problem list strings are

caused by an initial bulk load of historical information when a new location joins the data consortium. Figure 4 shows that, for user-entered observation values, the mapping percentage has been drifting downward slowly over time. Our hypothesis for this finding is that the initial observations we processed, such as blood pressure and weight, were fairly clean. Bringing in new types of observations, such as urine microalbumin, which have more variability in the format of possible values has led to an overall decrease in the percentage of values successfully resolved at the time of import.

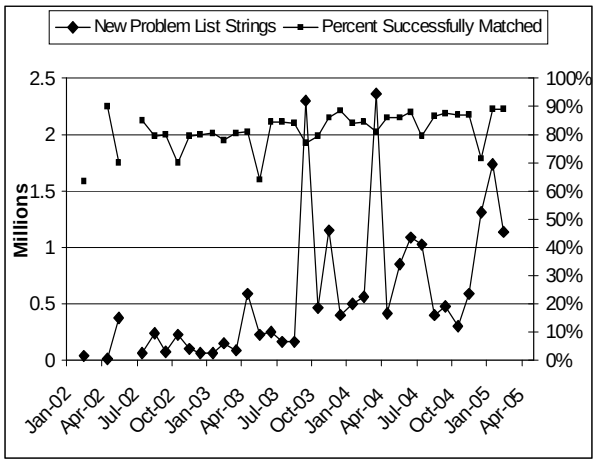


Figure 3: New problem list strings by month.

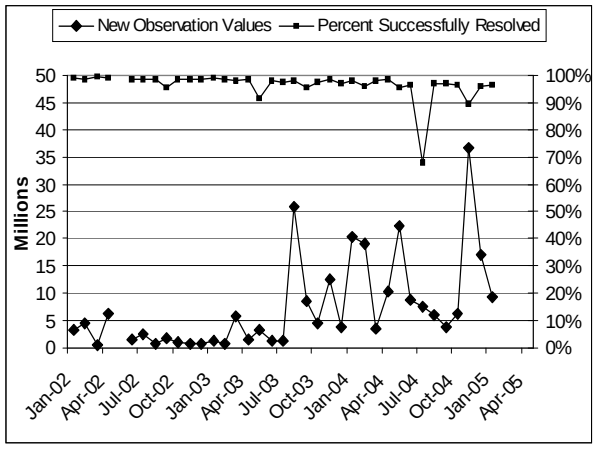


Figure 4: New observation values by month.

To keep both the mapping database and the mapping effort from growing without bounds, only user-entered problem list strings that are appear frequently (at least ~20 times) are mapped to a controlled term. Table 4 shows that although only 10% of *unique* problem list strings are mapped to controlled terms, this small percentage covers almost 86% of *all* problem list entries used by clinicians who submit their data to the national repository. The remaining 14% represent problem list strings that appear too infrequently to warrant mapping to a controlled term but are also stored in the repository.

Table 4: Strings to Concepts Mappings (06-Mar-05).

Total unique problem strings	1,926,359
Total unique problem strings mapped to concepts	206.189
% Strings mapped to concepts	10.7%
Total problem list entries in data warehouse	21,724,286
Total problem list entries is data warehouse mapped to concepts	18,563,867
% Entries mapped to concepts	85.5%

On average, the entire mapping effort consumes approximately 64 person-hours per month. For just user-entered problem list strings, approximately 612 new strings-to-concept mappings and 63 new problem list concepts are created monthly. When new locations join the data consortium, the import of historical data creates a large spike in unmapped terms requiring new mappings. Both the new-location start-up spikes and the growing baseline efforts are a continuous, never-ending commitment required to create and maintain a multi-institutional shared data repository with a controlled master concept terminology. Figure 3 and Figure 4 imply that widespread adoption of electronic medical records systems will not result in useful electronic data repositories without substantial on-going manual efforts to harmonize data.

Discussion

While the primary usage of EMR is to deliver clinical information and decision support to providers engaged in direct patient care, the ultimate benefit of electronic record adoption will be the ability to collect and analyze structured clinical data across patient populations to improve the practice of medicine for all patients. Without electronic medical record systems and a shared infrastructure for the aggregation of records across many care delivery sites, the goal of using information technology to improve quality, safety, or efficiency as envisioned by the National Health Information Network will not be realized. Missing from these discussions is insight into the substantial amount of infrastructure and work required to ensure that data used to create large aggregated data warehouses are accurate and comparable.

Creators of EMR systems must balance the strict use of controlled medical terminologies against end-user acceptance. Numerous studies have illustrated user reluctance to limit documentation to pre-defined controlled terms [5-7]. Thus all EMR systems in use today and for the near-term will contain a mixture of controlled terms and end-user variations, alterations, and completely new free-text strings. Automated and manual processing of these terms is required before these data can be entered into a warehouse. As the number of contributing systems and the breadth of clinical data contained within systems grows, the effort to ensure that only well-characterized and comparable data enters the data warehouse grows [8].

The MQIC consortium is extremely atypical from the envisioned scenario of ambulatory practice data warehouses. In this setting, data are extracted from the same commercial application and all locations have access to a single master set of terms that is managed centrally. Both features would not be present in the general case. In this respect, the data presented here represents the most optimistic estimate, a best-case lower bound, of the amount of work required to harmonize data from different practice locations. Despite the existence of national controlled terminologies, differences across EMR products are likely to increase the effort required to harmonize data substantially compared to the experience illustrated in Figure 3 and Figure 4. It is critical that discussions regarding the use of EMR data for aggregated data analysis recognize the significant effort required to create comparable data.

Summary

A shared technology infrastructure has been described that has enabled the Medical Quality Improvement Consortium to aggregate over 3.6 million detailed clinical records in an electronic data warehouse optimized for research and practice improvement. The architecture provides a model for at least one component of the proposed National Health Information Network. However, significant attention to the issues and efforts required to combine data collected from disparate locations which use differing technologies, terminologies and data entry methodologies must be included in the current discourse on the potential promises of wide-spread adoption of EMRs.

Acknowledgments

The authors gratefully acknowledge the efforts and contributions of Jeff Andersen, Mary Bowman, David Gonzalez MD, Tina Ho, Stuart Lopez, Sunil Luhadia, Teresa Smith, and Kevin Tabb MD.

Address for Correspondence
Thomas N. Ricciardi, PhD, Clinical Data Services Manager, GE Healthcare, 20540 NE Evergreen Parkway, Hillsboro OR 97124.
EMAIL: Tom.Ricciardi@med.ge.com

References

1. Institute of Medicine (U.S.). Committee on Improving the Patient Record., Dick RS, Steen EB. *The computer-based patient record: an essential technology for health care, revised edition.* Washington, D.C.: National Academy Press; 1997.

2. Institute of Medicine (U.S.). Committee on Quality of Health Care in America. *Crossing the quality chasm: a new health system for the 21st century*. Washington, D.C.: National Academy Press; 2001.

3. Thompson TG, Brailer DJ. The decade of health information technology: Delivering consumer-centric and information-rich health care. Framework for strategic action. 21 July 2004; http://www.hhs.gov/healthit/frameworkchapters.html . Accessed 10 March 2005, 2005.

4. Lieberman MI, Ricciardi TN, Spackman KA, Masarie FE. The use of SNOMED© CT improves querying of a clinical data warehouse. *Medinfo.* 2004;2004(CD):1721.

5. Embi PJ, Yackel TR, Logan JR, Bowen JL, Cooney TG, Gorman PN. Impacts of computerized physician documentation in a teaching hospital: perceptions of faculty and resident physicians. *J Am Med Inform Assoc.* Jul-Aug 2004;11(4):300-309.

6. Luff P, Heath C, Greatbatch D. Tasks-in-interaction: Paper and screen-based documentation in collaborative activity. Paper presented at: Proceddings of Computer Supported Cooperative Work, 1992.

7. Weir CR, Hurdle JF, Felgar MA, Hoffman JM, Roth B, Nebeker JR. Direct text entry in electronic progress notes. An evaluation of input errors. *Methods Inf Med.* 2003;42(1):61-67.

8. Marrs KA, Steib SA, Abrams CA, Kahn MG. Unifying heterogeneous distributed clinical data in a relational database. *Proc Annu Symp Comput Appl Med Care.* 1993:644-648.