Concept Dictionary Creation and Maintenance Under Resource Constraints: Lessons from the AMPATH Medical Record System

Martin C. Were, MD[1,2], Burke W. Mamlin, MD[1,2], William M. Tierney, MD[1,2], Ben Wolfe[1] and Paul G. Biondich, MD, MS[1,2]
[1]Regenstrief Institute, Inc. and [2]Indiana University School of Medicine, Indianapolis, IN.

Abstract
The challenges of creating and maintaining concept dictionaries are compounded in resource-limited settings. Approaches to alleviate this burden need to be based on information derived in these settings. We created a concept dictionary and evaluated new concept proposals for an open source EMR in a resource-limited setting. Overall, 87% of the concepts in the initial dictionary were used. There were 5137 new concepts proposed, with 77% of these proposed only once. Further characterization of new concept proposals revealed that 41% were due to deficiency in the existing dictionary, and 19% were synonyms to existing concepts. 25% of the requests contained misspellings, 41% were complex terms, and 17% were ambiguous. Given the resource-intensive nature of dictionary creation and maintenance, there should be considerations for centralizing the concept dictionary service, using standards, prioritizing concept proposals, and redesigning the user-interface to reduce this burden in settings with limited resources.

Introduction
A central feature of many enterprise-quality electronic medical record (EMR) systems is a dictionary-driven database model. Creating and maintaining dictionaries with desirable properties[1] has been a long-standing challenge for both developers and implementers of Electronic Medical Record (EMR) systems. It is an undertaking which requires subject matter expertise, and a significant financial and human-resource commitment. For this reason, growing numbers of vendors and large organizations are developing their own enterprise-wide terminology centers.[2,3] However, these approaches are not feasible within resource-constrained environments. For such organizations, strategies and tools specifically aimed at alleviating the burden of creating well-formed dictionaries and ensuring their 'graceful evolution'[1] are needed.

Three trends in medical informatics will increase the number of organizations with limited resources who have to deal with maintaining concept dictionaries. The first is the increasing adoption of EMRs in developing countries,[4] and the second is the adoption in small practices in the developed world[5] — because of limited human and financial resources, these places have a lot to gain from EMRs, but the same deficiencies make it difficult for them to manage a concept dictionary. The third trend is the growing role of open source software in medical informatics.[6] When compared with commercial software, open source software is less likely to come with embedded vocabulary content, the support infrastructure, or dictionary maintenance tools. Open source software's low up-front costs make them attractive to institutions with limited-resources, further compounding the challenge of creating and maintaining the concept dictionary.

The medical informatics literature has not comprehensively dealt with approaches to help alleviate the burden of creating and maintaining concept dictionaries in resource-limited settings. Some of the questions that need to be addressed in this area include: (1) whether to use standard vocabularies, like SNOMED and ICD-9 in these settings — the answer to this is not very clear, as these standards perform variably in different settings,[7,8] do not contain local terms, and can be costly; (2) whether it is easier, cheaper, and more practical to create a central terminology service for various implementers as opposed to having vocabulary managed at each institution; (3) how best to prioritize suggested concept proposals to deal with the more important ones; and (4) how to optimally design a user-interface to accommodate differences in workflow and decrease the number of new concept proposals that result from misspellings, lexical variations, and synonyms to concepts already in the dictionary — this becomes particularly relevant in settings where providers are not interacting with the computer directly (e.g. settings where data-entry clerks, who usually have limited medical knowledge, input the data into the computer).

Approaches to these issues should be guided by lessons and demands on the ground. As a step in this direction, we describe the creation and evaluation of a concept dictionary and subsequent concept proposals for an open source EMR implemented in the resource-limited setting of Western Kenya. In this evaluation, we determine: (1) how well the initial

concept dictionary met the needs in the local setting; (2) the characteristics of new concept proposal requests; and (3) the relationship between these characteristics and the number of times a new concept was requested.

Methods

Setting

We conducted this evaluation at the Academic Model for Prevention and Treatment of HIV-AIDS (AMPATH) program[9] in Western Kenya. AMPATH is a collaborative initiative between Indiana University, U.S.A. and Moi University, Kenya which provides comprehensive care for patients infected with HIV. The program was initiated in 2001 and currently takes care of >40,000 HIV+ patients who have made >400,000 visits to 20 outpatient clinics.

AMPATH started using the AMPATH Medical Record System (AMRS) on Feb-15-2006 as its sole medical record. AMRS is an implementation of OpenMRS (http://openmrs.org), an open source EMR currently being widely deployed in Eastern and Southern Africa.[10] All data in AMRS are stored as coded concepts to allow for easy retrieval and analysis. The concept dictionary is thus at the core of the EMR, and includes terms for diagnoses, tests, procedures, drugs, and other general terms for questions and potential answers.

We evaluated the initial concept dictionary terms which were diagnoses, findings and symptoms. We also analyzed new concept proposals for these categories during the first year of AMRS use, between Feb-15-2006 and Feb-14-2007. During this time, 237 providers recorded >9 million discrete clinical observations that were entered into AMRS by 28 data-entry clerks.

Creating the Initial Concept Dictionary

Two physicians, Drs. Paul G. Biondich (PGB) and Terry J. Hannan (TJH) created the concept dictionary for the initial installation using principles from Cimino's desiderata[1], the ISO specifications,[11] and Chute's framework[12]. Both PGB and TJH have extensive experience in developing and implementing medical information systems. Working over a two year period, they compiled dictionary terms based on: (1) concepts used on AMPATH encounter forms, (2) the minimum dataset for HIV care in developing countries,[13] (3) suggestions by AMPATH providers, and (4) the pre-existing concepts in the Mosoriot Medical Record system,[14] an MS-Access®-based system previously used at one of the AMPATH clinics. This dictionary not only contained base concepts, but additionally described a rich synonym list, based upon reported local naming conventions.

Proposing New Concepts

When providers take care of a patient at an AMPATH clinic, they do not enter data into AMRS directly, but rather complete paper-encounter forms which contain questions and answers that have already been encoded into the AMRS concept dictionary. In some cases they are allowed to write down other answers as free text if these are not included as coded choices on the encounter form. Information on the encounter form is entered into AMRS by data-entry clerks who have little medical knowledge. When these clerks enter a term from the paper form, the system dynamically generates a list of all concepts that contain exact string matches to the letters already typed (Fig. 1).

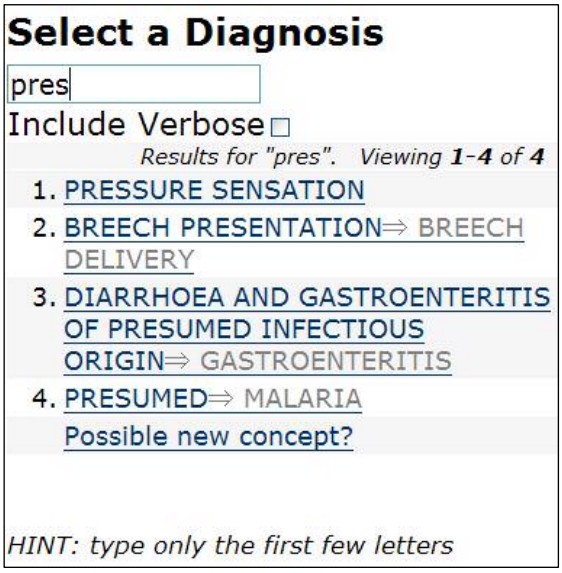

Fig. 1 Example for searching the concept dictionary terms in AMRS.

The only way a clerk can enter a datum for which there is no dictionary term is to select '**Possible new concept?**' and type it in (Fig. 2). A concept is thus proposed as many times as it appears on the encounter forms. The proposals are stored in the encounter data as placeholders and copied into a queue. Once the proposals have been moderated, the placeholders are replaced with appropriate coded values.

Analyses of Initial Concept Dictionary

We determined the number of concepts (diagnoses, finding or symptoms) which were used at least once over the evaluation period. We also determined the percentage of all answers provided on encounter forms which used these concepts.

Analysis of New Concept Proposals

To allow for tabulation and analysis of new concept proposals, we chose not to add any of them into the

initial concept dictionary during the 12-month period of this evaluation. This did not affect patient care as all encounter forms were stored in the patient's paper chart and these were available to the providers.

Fig. 2: Example for submitting a new concept proposal in AMRS.

The number of times each new concept was requested was determined. We then randomly sampled 20% of the new concept proposals, which were then characterized by one of the researchers, Martin C. Were, into categories modified from Wang et al.[15] and Robinson et al.[16] The categories included:

- Dictionary Deficiency: The new concept proposal did not exist in the concept dictionary.
- Synonym: The new concept proposal was either a true synonym of a term in the dictionary (e.g. 'Boil' as a synonym of 'Furunculosis'), an impure synonym (e.g. 'Enteritis' classified as a synonym of 'Gastroenteritis'), or a lexical variant (e.g. 'asthmatic' classified as synonym of 'asthma').
- Misspelling.
- Ambiguous term: The term entered did not make much sense (e.g. 'Foot root'), was an uncommon term specifically used at that site (e.g. 'FGC'), or had multiple possible meanings (e.g. 'Adherence').
- User Interface Failure: The new concept existed in the dictionary but the clerk still suggested it as a new concept proposal.
- Complex term: The new concept proposal included more than one potential concept (e.g. 'Ascites/Bronchitis'), or had a modifier (e.g. 'Severe Anemia' and 'Right sided lung collapse').

A new concept could contain more than one of the characteristics above – for example, the concept proposal 'HEMORHOIDS' had a misspelling, and also represented a dictionary deficiency as the initial dictionary lacked the term 'HEMORRHOIDS'.

We also looked for trends in the characteristics of new concept requests as a function of how many times they were proposed. This was done for all concepts proposed more than once.

Results

Analysis of Initial Concept Dictionary Use

The initial dictionary contained 424 concepts for clinical findings, symptoms, and diagnoses. Out of these, 380 (87%) were used at least once as answers on encounter forms. These concepts accounted for 79,336 of the 90,510 (88%) answers given on encounter forms over the study period. The other 11,174 answers were not in the concept dictionary, requiring the data-entry clerk to make a new concept request.

Analysis of New Concept Proposals

Over the study period, a total of 5,137 unique new concepts were requested. Out of these, 3,954 (77%) were requested one time, 484 (9%) twice, 209 (4%) three times, and the other 490 (10%) requested more than three times. 12 (0.2%) of the concepts were proposed over 100 times, with 'ENTERITIS' having the most requests at 461.

Table 1. Analysis of New Concept Proposals by various characteristics*

Characteristics	N	%
Dictionary Deficiency	418	41
Synonym	199	19
Misspelling	252	25
Complex Term	420	41
Ambiguous Term	179	17
User Interface Failure	10	1

* A particular new concept proposal could have one or more of these characteristics.

Table 1 shows the characteristics of a randomly selected sample of 1027 (20%) new concept proposals. Of the 199 terms judged as synonyms, 78 (39%) represented true synonyms, 89 (45%) impure synonyms, and 32 (16%) lexical variants. Out of the 420 complex terms, 93 (22%) were proposals which contained more than one concept, 149 (36%) had a location modifier (e.g. 'Facial rash' or 'Right-sided pneumonia'), 84 (20%) had a quality modifier (e.g. 'mild anemia'), and 94 had other types of modifiers (e.g.

'r/o pneumonia'). Only 33 terms (18% of ambiguous terms and 3% of all terms) were classified as being site-specific.

Fig. 3 shows the frequency of various characteristics as a function of how many times concepts were requested.

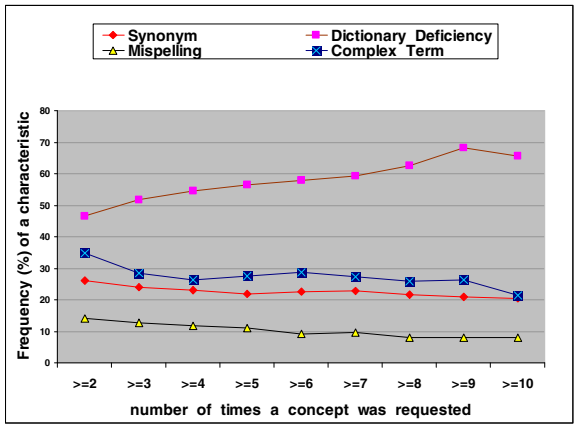

Fig 3: Relationship between the characteristics of concept proposals and the number of times a concept was requested.

Discussion

Creating and maintaining a well-structured concept dictionary is challenging. It is made even more difficult in settings of limited human and financial resources. In this paper, we report our experience and the lessons we have learned during the first year of using the AMRS system.

We show the benefit of investing time to create the initial dictionary. 380 concepts in the initial dictionary provided 88% of all the answers used over the first year. This was achieved after two years of hard work by experts in this area. Such a time-intensive approach might not be feasible in many settings and alternatives include: (1) implementers concentrating on concepts they think will be absolutely necessary for their practice (like those used on encounter forms), and adding any new concepts as the need arises; (2) Using standard vocabularies such as ICD-9 - the utility of these in resource-poor settings still needs formal evaluation; and (3) a consideration by implementers and developers who have already created dictionaries to share them with others (we provide a version of the OpenMRS dictionary at http://demo.openmrs.org/openmrs/dictionary/). Ideally, concept dictionaries put in the public domain should have links to standard vocabularies like ICD-9 and LOINC to make their adoption easier for those not familiar with a particular EMR implementation.

The AMRS concept dictionary was built on the premise that it was impossible to come up with all the concepts needed at the outset, but that the local community could help augment the dictionary. The 418 new proposals which revealed a dictionary deficiency support the feasibility of this approach. We also observed that the more frequently a concept was proposed, the more likely it was to represent a deficiency in the dictionary (Fig 3). This suggests that under resource constraints, time and resources could be better spent working on concepts that have been proposed more frequently. It seems to us that it is a good design principle to keep track of how many times a particular concept is requested as this can also help in deciding the level of granularity needed for a concept.

A significant number of the proposed concepts were complex terms. From these were learned that giving the providers some basic understanding of how data is encoded into the system could help reduce the numbers of complex terms suggested. Furthermore, our observation that there are a significant number of modifiers used point to the importance of addressing issues around pre- and post-coordination of concepts early during the development and implementation of a system. Advanced EMR developers might find it valuable to incorporate automated term dissection[17] and composition[18] algorithms into their software to help deal with complex terms.

The requested concepts also inform us on how user-interfaces should be designed when the person entering the data into the computer has little medical experience. Misspellings, which are usually off by one or two letters, and simple lexical variants can be matched correctly if approximate string comparator algorithms, like the Levenshtein Edit distance algorithm[19], were incorporated into the system. Furthermore, it would be important to understand the social context within which the clerks operate, the role of memory, and the decision-making strategies they employ when proposing new concepts.[20]

Finally, we observe that site-specific terms only made up a small percentage of all terms requested (3%). This finding is encouraging, since, in resource-limited settings, it might be preferable for several implementations to use a centralized vocabulary service. Such a service would: (1) facilitate data-sharing, which is critical for the success of such collaborative research projects like those being conducted by the International Epidemiological Databases to Evaluate AIDS (IeDEA),[21] (2) reduce the number of terms that have

slightly different meanings, thus making data easier to extract from the system, and (3) reduce the burden of creating the initial concept dictionary.

It is our hope that the lessons from our experience will be valuable for developers and implementers of EMRs in general, and provide specific benefit to those in resource-limited settings.

Acknowledgements
This work was performed at the Regenstrief Institute, Indianapolis, IN, and was supported by grant T15 LM07117 from the National Libarry of Medicine. Special thanks to Dan Vreeman, Mike Weiner, Marc Overhage, and Shaun Grannis for their advice.

References

1. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf. Med. 1998; 37:394-403.
2. Dolin RH, Mattison JE, Cohn S, Campbell KE, Wiesenthal AM, et al. Kaiser Permanente's Convergent Medical Terminology. MedInfo 2004; 346-350.
3. Lincoln MJ, Brown SH, Nguyen V, Cromwell T, Carter J, Erlbaum M, Tuttle M. U.S. Department of Veterans Affairs Enterprise Reference Terminology Strategic Overview. MedInfo 2004; 391-395.
4. Fraser HS, Biondich P, Moodley D, Choi S, Mamlin BW, Szolovits P. Implementing electronic medical record systems in developing countries. Informatics in Primary Care 2005; 12: 83-95.
5. Gans D, Kralewski J, Hammons T, Dowd B. Medical groups' adoption of electronic health records and information systems. Health Affairs 2005; 25(5): 1323-1333.
6. Hogarth MA, Turner S. A study of clinically related open source software projects. AMIA Annu Symp Proc. 2005; 330-334.
7. Payne TH, Murphy GR, Salazar. How well does ICD9 represent phrases used in the medical record problem list. Proc. Annu.Comput Appl Med Care. 1992; 654-657.
8. Dykes PC, Currie LM, Cimino JJ. Adequacy of evolving national standardized terminologies for interdisciplinary coded concepts in an automated clinical pathway. J Biomed Inform. 2003; 36(4-5): 313-25.
9. Voelker, R. Conquering HIV and Stigma in Kenya. JAMA. 2004; 292: 157-159.
10. Mamlin B, Biondich PG, Wolfe BA, Fraser HS, Jazayeri D, et al. Cooking up an open-source EMR for developing countries: OpenMRS – a recipe for successful collaboration. AMIA Annu Symp Proc 2006; 529-33.
11. ISO/TS 17117:2002(E): Health informatics-controlled health terminology-structure and high-level indicators: technical committee ISO/TC 215, Health Informatics; 2002.
12. 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; 5:503-10.
13. Tierney WM, Beck EJ, Gardner RM, Musick B, Shields M. Viewpoint: a pragmatic approach to constructing a minimum data set for care of patients with HIV in developing countries. J Am Med Inform Assoc. 2006; 13: 253-260**.**
14. Rotich JK, Hannan TJ, Smith FE, Bii J, Odero WW, et al. Installing and implementing a computer-based patient record system in sub-saharan Africa: the Mosoriot Medical Record System. J Am Med Inform Assoc. 2003; 10:295-303.
15. Wang SJ, Bates DW, Chueh HC, Karson AS, Maviglia SM, et al. Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool. Int Journal of Med Info. 2003; 72:17-28.
16. Robinson D, Schulz E, Brown P, Price C. Updating the Read Codes: User-interactive Maintenance of a Dynamic Clinical Vocabulary. J AM Med Inform Assoc. 1997; 4:465-472.
17. Elkin PL, Bailey KR, Ogren PV, Bauer BA, Chute CG. A randomized controlled trial of automated term composition. AMIA Annu Symp Proc. 1998; 765-769.
18. Elkin PL, Bailey KR, Ogren PV, Bauer BA, Chute CG. A randomized double-blind controlled trial of automated term dissection. AMIA Annu Symp Proc. 1999; 62-66.
19. Levenshtein Edit distance (Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady. 1966; 10(8):707-10.
20. Patel VL, Arocha JF, Kaufman DR. A Primer on Aspects of Cognition for Medical Informatics. J Am Med Inform Assoc. 2001; 8:324-343.
21. International epidemiological Databases to Evaluate AIDS. Available at: http://www.iedea-hiv.org/. Accessed Jul 17, 2007.