Linking Information Systems for HIV Care and Research in Kenya

Alicia F. Guidry
University of Washington
Box 357240
Seattle, WA, USA 98195-7240
+1 206 395 4324
aguid@uw.edu

Judd L. Walson
University of Washington
Box 359909
Seattle, WA, USA 98195-9909
+1 206 543 4278
walson@uw.edu

Neil F. Abernethy
University of Washington
Box 357240
Seattle, WA, USA 98195-7240
+1 206 616 2813
neila@uw.edu

ABSTRACT
The provision of HIV care in developing countries may involve complex and overlapping resources; including government-run facilities non-governmental organization (NGO) or international non-governmental organization (INGO) supported services and research affiliated clinics. These resources are often motivated and funded by distinct health priorities and as a result, standards for clinical data representation and exchange are rare and data management systems are often redundant. Open-source systems such as OpenMRS and OpenClinica provide an opportunity to leverage available systems to improve standards and increase interoperability. Nevertheless, continuity of care and data sharing between these systems remains a challenge, particularly in populations with changing health needs, high mobility, and inconsistent access to health resources. As a prerequisite to improving interoperability between systems, use cases for clinical information exchange are first identified. We then characterize data models from nine clinical information systems, standards, and ontologies pertinent to HIV clinical care and research in Kenya. The data fields commonly used as patient identifiers are summarized, including name, date of birth, family relations and location. Finally, we present a prototype ontology to describe data standards and to enable mapping between data elements in diverse information systems.

Categories and Subject Descriptors
I.2.4 Knowledge Representation Formalisms and Methods

D.2.12 Interoperability

General Terms
Algorithms, Management, Performance, Reliability, Standardization, Languages, Verification.

Keywords
Interoperability, data standards, HIV, electronic medical records, ontologies, data integration, open source software

IHI'10, November 11–12, 2010, Arlington, Virginia, USA.

1. INTRODUCTION
The health care landscape in developing countries includes a myriad of different avenues by which a patient can receive medical care. These avenues include but are not limited to NGO-based clinics, external support from agencies such as United States Agency for International Development (USAID), President's Emergency Plan for AIDS Relief (PEPFAR), Centers for Disease Control and Prevention (CDC), and World Health Organization (WHO), and local Ministries of Health. Many such efforts are focused on individual diseases or conditions, particular populations, or fixed regions. They also draw from distinct funding sources and have distinct monitoring and evaluation (M&E) requirements. Accordingly, each agency or clinic typically collects its own data and, if the site is computerized, has their own associated information system. This approach while perhaps effective for any single particular system, does not benefit the health care system as a whole. In addition, such systems may fail when a patient moves or requires overlapping care from multiple systems.

These constraints often result in health care information becoming siloed within a single provider system, resulting in unnecessary redundancy of effort and resources, duplicate data entry, fragmented and unsharable health records, poor allocation of time for both patients and clinicians. While information systems alone cannot improve access to care, data standards and system interoperability may mitigate the impact of some of these issues.

Because of the disparate purposes, designs, and architectures of information systems, data integration can be difficult to achieve in this setting. Furthermore, system developers are confronted by an array of legacy functional needs, clinical aims (treatment, prevention, and clinical trials), and both official and ad hoc standards. To address this diversity of both information uses and data models, we highlight possible use cases for querying across clinical information systems (CIS). We next describe the current level of standardization of patient identifiers needed to link data between systems. Finally, we introduce a prototype patient identifier ontology, which can be used to aid in the data integration of heterogeneous CIS.

Establishing standards in an environment with existing competing data models can be difficult. It is important first to establish the level of agreement between data representations and the overlap in domain coverage. In a prior study of identifying information used in tuberculosis contact investigation forms in 50 states and 3 countries, Abernethy et. al. [1] revealed a broad range of data field frequency used to identify patient contacts. In the setting of fragmented HIV care systems and competing guidelines, we must

assess the current state of standardization particular to this context.

Once the scenarios of data integration are identified and the level of de facto standardization among patient identifiers has been established, our next step will be to develop a platform to enable data linkage between systems. A formal ontology defines the concepts and semantics associated with a particular domain [2]; this data model will enable complex and flexible mappings between semantically related identifiers such as "location" and "address". Ontologies are often used to facilitate data integration between heterogeneous systems. Additionally, adoption of ontologies for use in developing information systems can encourage implementers to conform to the data standards set forth by the ontology. In the healthcare domain, the most important component of data integration is identifying the patient whose data needs to be viewed. Unfortunately, this can be problematic when the need to query across systems arises.

2. DATA INTEGRATION USE CASES

Data integration has many different uses, specifically within the medical field. We describe four medical situations: continuity of care, querying across systems, scheduling and wasted resources that would benefit from data integration of CIS.

Provide Continuity of Care

Sharing information for the optimal care of patients is often a challenge if the population is mobile, if treatment is sought opportunistically from among the few available resources, or if patients seek care from different sources for distinct conditions. In the best scenario, a patient might present with a paper record from another clinic summarizing their treatment or vaccination history. The ability to transfer patient data between currently siloed systems could avoid complications such as misdiagnosis or contraindicated therapy. (For example, a paper-based transfer record could omit a patient's allergy to Septrin, resulting in another provider initiating Septrin prophylaxis, precipitating a severe reaction).

Query Across Systems

Patients that exist in different CIS are often compiled to create cohorts for clinical trials. Clinicians determine the inclusion criteria for their trial, and later query systems to determine patients who will include or excluded from their trial. Another example of this use case would be querying other providers' systems to identify prior medical records for a given patient.

Harmonize Scheduling

Patients seeking routine care at a local district hospital HIV clinic might also enroll in an anti-retroviral treatment (ART) clinical trial being run out of the same clinic. However, the visits required by routine treatment and study protocols may differ. Electronic medical record systems (EMR) and the clinical trial management system are typically are not connected, making it is impossible to synchronize visits in the systems. Aligning schedules to make treatment viable and provide reminders for patients will prevent protocol and treatment lapses.

Prevent duplication of effort

Patients may receive blood draws for CD4 counts or viral loads as part of routine care, and again for clinical trial protocols. However, availability of a recent CD4 count from a clinical trial would be sufficient for clinical care.

The use cases presented above are common throughout the medical community. However, in some areas it is often unavoidable, especially in clinical care settings that lack EMR, have heterogeneous CIS or lack interoperability between clinical sites.

3. METHODS

A convenience sample of nine data models, information systems, and standards utilizing patient identifiers were obtained: the RadLex [3] ontology, the OpenMRS [4] and OpenClinica [5] database schemas, a peer-reviewed paper on the Mosoriot Medical Records System (MMRS) [6] and a report by RAND Health on Unique Patient Identifiers [7], the Kenya National AIDS/STD Control Programme Comprehensive Care Patient Card (NASCOP) Blue Card [8], Centers for Disease Control and Prevention – Council of State and Territorial Epidemiologists (CDC-CSTE) [9] and WHO patient monitoring guidelines[10], and the Johns Hopkins Patient Identification System website [11].

Technical information on data models were obtained from online manuals, published descriptions, reference websites, and personal communication with system developers. In order to include an example ontology in the sample, we searched the BioPortal [12] website which facilitates the sharing of ontologies in the biomedical community. A search using the keywords "patient identifier" returned three ontologies: RadLex, RadLex in OWL (Web Ontology Language) and Logical Observation Identifier Names and Codes (LOINC) [13]. LOINC was not included in this study due to its specialized focus on laboratory results. RadLex and RadLex in OWL's patient identifying information were identical, hence they were treated as a single source for our purposes. RadLex provided six patient identifier fields. This information can be seen in Table 1, following this discussion.

In Kenya, patients enrolled in government health care clinics have data recorded on a card provided through the NASCOP, an agency of the Kenya Ministry of Health. The so-called "Blue Card" standardizes the demographic and treatment information collected on each patient. Patients bring the card to each clinic visit, during which the card is updated with recent information. However, the majority of health care facilities in Kenya do not have a clinical information system and therefore most of this information is either kept on paper forms (internal clinical forms) or the blue card.

The NASCOP Blue Card is often compared to the WHO patient monitoring guidelines for HIV. This dataset is used throughout the world for the treatment of HIV/AIDS and is a part of the WHO standard for electronic medical record systems. In addition to the WHO standard system, the Centers for Disease Control have created the CDC-CSTE dataset to aid in the standardization of information requested by CDC and other public health agencies related to disease surveillance.

The peer-reviewed paper (MMRS) and report (RAND) were found using a Google Scholar search for patient identifier information. The MMRS paper specifically discusses the patient identifier information needs related to implementing an EMR in Kenya. Information about what patient information would be needed to implement a unique patient identifier for the United States is discussed in the RAND report. Finally, the Johns Hopkins Website discusses their Patient Identification System. This system assigns medical record numbers to patients based on

their personal identifying information, and facilitates merging the medical records of patients having duplicate record numbers.

- RadLex - Ontology; used by Radiologists
- NASCOP Blue Card - Patient card; used throughout Kenya by all HIV/AIDS patients
- WHO Dataset - Recommendations by the WHO for HIV/AIDS care; used through out the world
- CDC-CSTE Dataset - Recommendation by the CDC-CSTE Working Group; used to standardize infectious disease investigation
- OpenClinica - Database schema; used to capture clinical research data
- OpenMRS - Database schema; used to capture observations from clinic visits
- MMRS Paper - describes EMR system; used in Kenya
- RAND Report - describes the procedures used to identify patients;
- Johns Hopkins Website - describes a Patient Identification System; used by Johns Hopkins Medicine to identify patients in disparate clinical information systems

Composition of the data models was tabulated in a Microsoft Excel spreadsheet (Table 1). The rows of the spreadsheet are the variable or column names from the article, database schema or ontology. The columns of the spreadsheet were the names of the system or ontology described. If a system or ontology used a particular variable or column name to identify patients an 'x' was placed in the box corresponding to that column and row. This resulted in 35 data fields, of which 34 were distinct. Data fields were organized according to their semantic content into the following categories: Numeric identifiers, Patient information, Relative information, and Location.

Data from the spreadsheet was used to create a patient identifier ontology using the Protégé 4.1 [14] system (Figure 1). We created an OWL-based ontology having 46 **classes** and 32 **properties** using the bottom-up ontology development approach, starting with concepts from the data model summary as seeds. The number of classes outnumbers the source fields due to generalization of the data model to accommodate broader data types and instances. For example, *Clinical* study is a type (subclass) of *Research study* to allow for the possibility of non-clinical studies such as the behavioral surveys important for HIV prevention. Similarly, *Patient* and *Provider* both generalize to the class *Person*, each of whom will have some distinct and shared properties.

Properties (also known as slots) describe the features of class members. For example, each *Person* (including a *Patient*) has a *Gender* and a *Name*. The *hasGender* property links an instance of *Person* to an instance of the class Gender, such as in the assertion "(hasGender Tony Male)". Most properties in our ontology correspond directly to the data fields compiled in the rows of Table 1.

4. RESULTS

Data Model Analysis

Of the 34 unique data fields only *date of birth* appeared in all sources. However, use of this field may result in misclassification bias, as exact day, month and year of birth is unknown areas of the world [6]. Only *gender* and *first name* appeared in six of the nine sources. Five used *Patient middle name,* while *Patient last name*, *age*, *telephone* and *postal code* appeared in four systems. Among data categories, the most common numeric identifiers were Medical record number and Unique patient identifier. The most common patient information fields was *date of birth*, however all data models included some version of patient name. The most common location information fields were *Street address*, *Postal code*, and *Telephone*. Data fields pertaining to patient relatives were only used in 3 data models in our sample, and no field occurred more than once.

Comparing data models, the CDC-CSTE, OpenMRS, and WHO HIV/AIDS minimum dataset used similar collections of location data. Among patient information data fields, similarities are seen between the RadLex and NASCOP models (which depend heavily on *age* and *gender*), and between MMRS, OpenMRS and Rand Health (which specify *first name* and *middle name* with few other fields). This likely reflects the derivation of OpenMRS from the MMRS system. Other similarities exist between systems but are less easily grouped into meaningful sets.

Patient Identifier Ontology

Building on the results of the data model analysis, we have created a prototype patient identifier ontology to capture the explicit semantics of each data field in a formal, computable description. The class hierarchy of this ontology are seen in Figure 1. Two main divisions in the ontology are between *Physical* and *Conceptual Things,* a common distinction in knowledge representation. Entities such as patients, relatives, providers and organizations are *Physical Things* that have locations, addresses, and names. By contrast *Conceptual Things* describe things such as *Attributes* (e.g. *Gender*), *Identifiers* (e.g. *StudyID*), and *Roles* (e.g. *Patient* or *Provider*).

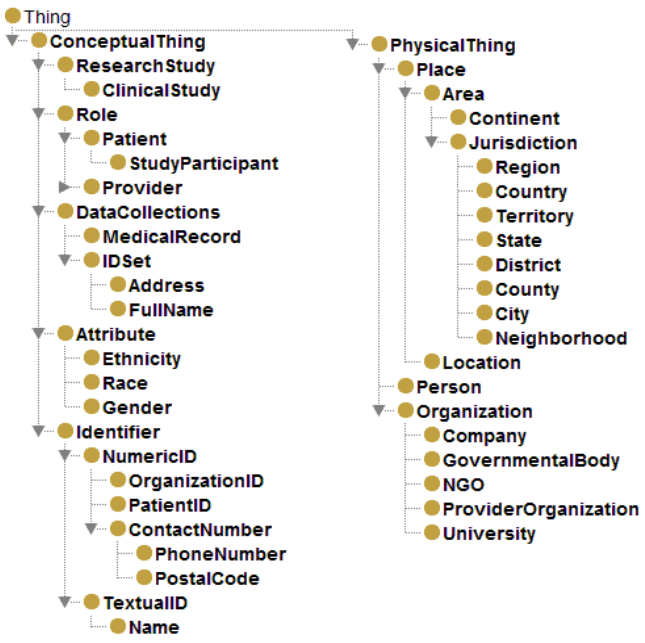

Figure 1. Patient Identifier Ontology

Table 1 is an abbreviated representation of the data model survey. Data fields were categorized into four types: Numeric IDs (identifiers), Patient information, Relative information (such as Next of Kin or Mother's maiden name), and Location information. Data fields only found in one information system (except Numeric IDs) were collapsed for clarity, hence 25 groupings of the 34 fields are shown. Relative information, for example, includes the fields *Next of Kin, Next of Kin Last Name, Next of Kin First Name, Next of Ken Telephone, Father's Name, Mother's Maiden Name* and *Patient's Mother's First name* (each of which was found in only one data model). Likewise *Location* subsumes the fields *Location, Sub-Location, Nearest Landmark and Nearest H/Facility,* all of which appeared solely in the NASCOP dataset, likely reflecting derivation from the rural setting in Kenya.

Category		RadLex	OpenMRS	OpenClinica	WHO	MMRS	NASCOP	CDC-CSTE	Johns Hopkins	RAND Health
Numeric IDs	**Medical Record Number**	x	x							
	ClinicID				x					
	StudyID			x						
	SSN (Last 4)									x
	Unique Patient Identifier				x		x			
Patient information	**Age**	x			x		x	x		
	Date of Birth	x	x	x	x	x	x	x	x	x
	Ethnicity	x						x		
	Race							x	x	
	Gender	x		x	x		x	x	x	
	Full Name	x	x				x			
	First Name		x		x	x		x	x	x
	Middle Name		x			x		x	x	x
	Last Name				x	x		x	x	x
	Marital Status				x		x		x	
Relative information	**(includes seven fields: Next of Kin, etc.)**					x		x	x	
Location information	**District**		x				x			
	Patient's Home Village		x			x				
	Location (four location fields)						x			
	Street Address		x		x		x	x		
	City		x		x			x		
	State/territory		x		x			x		
	Zip Code		x		x			x		x
	Country		x					x		
	Telephone				x		x	x	x	

5. CONCLUSIONS

We have provided a patient identifier ontology, which can be used to facilitate data integration and exchange. However, this ontology alone is not enough to provide this functionality. Along with mapping the ontology through some sort of manual, automatic or semi-automatic means, one needs to determine whether or not the information returned is accurate (i.e. the patient being returned is the patient being searched for).

Patient matching can be done many different ways. However, most important is the need to determine which information is required to identify a correct match. As we have shown systems can use any number of data points to identify a patient. The majority of sources we used identified Patient Date of Birth, Gender, Name, Age and Zip Code and Telephone number among other things to identify patients. However, when integrating disparate sources one should take into account that the information could be missing, recorded in a different format or misspelled.

One limitation of this study is the exclusion of clinical vocabularies or terminologies. Resources such as the Unified Medical Language System (UMLS) metathesaurus [15] and the National Cancer Institute (NCI) metathesaurus link several existing vocabularies that may contain several concepts relevant to this domain. Clinical vocabularies, while not initially created to serve the specialized needs of developing countries, are

nonetheless increasingly being used in this setting; hence their coverage of patient identifiers may become increasingly relevant. Similarly, our study did not include data exchange standards such as Health Level Seven (HL7) [16] or semantic web data models, both of which will conceivably be used with a higher frequency as collaborative sites begin to share more data. More generally, a broader search for relevant data models and ontologies would bolster our results. A more comprehensive analysis of systems and standards in use would improve our chances of realizing a standardized model.

We plan to expand our survey to other information systems used in the East Africa region. Furthermore, given that patient names are used as primary identifiers in this context, and that they are transliterated inconsistently into a Roman alphabet, we will assess the fidelity of name matching between systems using record linkage algorithms including string comparison and the Fellegi-Sunter algorithm.

Finally, we intend to evaluate and extend our ontology. We will first test the flexibility of the ontology by creating mappings to identifiers from samples of information systems that were and were not included in our bottom-up design approach. To reach our goal data exchange between CIS, we will also seek to include clinical data and information used in the care and treatment of HIV/AIDS. These extensions will help users of these and other information systems achieve data integration amidst a patchwork health care system and competing, duplicative, or incompatible patient identifiers and clinical data.

Kenya has recently begun investigating the standardization of medical data and creation of interoperable systems. The Kenya Bureau of Standards, the Ministry of Health and the International Training & Education Center for Health (I-TECH) are working with stakeholders to identify standards to be used throughout the country as they continue to build capacity for a nationwide e-Health Policy. While this initiative's focus is not to create or use ontologies, it does set out to provide a standard for the types of technologies that will be required for interoperability throughout the country. Our data model may inform these efforts.

In the United States clinical and research data are rarely pooled because of legal regulations. However it is not uncommon for secondary use of EMR data to occur after institutional review board approval. In developing countries this happens even less, primarily because of the lack of electronic medical data. However, one of the benefits of interoperable data is better reporting which can lead to increased funding, which can be a major motivator in acceptance.

6. ACKNOWLEDGMENTS

We would like to acknowledge the efforts the developers of the data models, systems and standards used in this study, which included open source systems. We also thank the staff of the Kenya Medical Research Institute/UW projects and the many clinic staff in Kisii Provincial Hospital, Kisumu District Hospital and Kilifi District Hospital for their input. AG was supported by NLM training grant 5T15LM007442-08.

7. REFERENCES

[1] Abernethy, N. *Automating social network models for tuberculosis contact investigation*. Stanford University, 2005.

[2] Chandrasekaran, B., Josephson, J. and Benjamins, V. What are ontologies, and why do we need them? *IEEE Intelligent Systems and their applications*, 14, 1 1999), 20-26.

[3] NCBO BioPortal: RadLex, http://bioportal.bioontology.org/ontologies/40885, accessed on May 2, 2010.

[4] OpenMRS - OpenMRS, http://openmrs.org/wiki/OpenMRS, accessed on June 3, 2010.

[5] Clinical Trial Software | OpenClinica, http://openclinica.org/, accessed on June 3, 2010.

[6] Hannan, T. J., Rotich, J. K., Odero, W. W., Menya, D., Esamai, F., Einterz, R. M., Sidle, J., Sidle, J., Smith, F. and Tierney, W. M. The Mosoriot medical record system: design and initial implementation of an outpatient electronic record system in rural Kenya. *International Journal of Medical Informatics*, 602000), 21-28.

[7] Hillestad, R., Bigelow, J. H., Chaudhry, B., Dreyer, P., Greenberg, M. D., Meili, R. C., Ridgely, M. S., Rothenberg, J. and Taylor, R. *Identity Crisis: An Examination of the Costs and Benefits of a Unique Patient Identifier for the U.S. Health Care System*. Rand Health, 2008.

[8] National AIDS/STD Control Programme (NASCOP), http://www.aidskenya.org/, accessed on May 2, 2010.

[9] Common Core Data Elements 2009 CSTE Position Statement 09-SI-01, http://cste.org/ps2009/09-SI-01.pdf, accessed on May 2, 2010.

[10] WHO *PATIENT MONITORING GUIDELINES FOR HIV CARE AND ANTIRETROVIRAL THERAPY (ART)*. http://who.int/3by5/capacity/ptmonguidelinesfinalv1.pdf, accessed on May 2, 2010.

[11] Information Technology @ Johns Hopkins-Patient Identification System, http://it.jhu.edu/fas/pid.html, accessed on June 3, 2010.

[12] NCBO BioPortal: Welcome to the NCBO BioPortal, http://bioportal.bioontology.org/, accessed on May 2, 2010.

[13] Logical Observation Identifiers Names and Codes (LOINC®) — LOINC, http://loinc.org/, accessed on May 2, 2010.

[14] The Protege Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu/, accessed on May 2, 2010.

[15] Unified Medical Language System (UMLS) - Home, http://www.nlm.nih.gov/research/umls/, accessed on June 3, 2010.

[16] Health Level Seven International - Homepage, http://www.hl7.org/, accessed on June 3, 2010.