Evaluation of family history information within clinical documents and adequacy of HL7 clinical statement and clinical genomics family history models for its representation: a case report

Genevieve B Melton,[1,2] Nandhini Raman,[1] Elizabeth S Chen,[3] Indra Neil Sarkar,[3] Serguei Pakhomov,[1,4] Robert D Madoff[2]

[1]Institute for Health Informatics, University of Minnesota, Minneapolis, Minnesota, USA
[2]Department of Surgery, University of Minnesota, Minneapolis, Minnesota, USA
[3]Center for Clinical and Translational Science, University of Vermont, Burlington, Vermont, USA
[4]Department of Pharmaceutical Care & Health Systems, University of Minnesota, Minneapolis, MN

Correspondence to
Dr Genevieve B Melton, Department of Surgery and Institute for Health Informatics, University of Minnesota, 420 Delaware Street SE, MMC 450, Minneapolis, Minnesota, USA; gmelton@umn.edu

Received 3 August 2009
Accepted 31 January 2010

ABSTRACT

Family history information has emerged as an increasingly important tool for clinical care and research. While recent standards provide for structured entry of family history, many clinicians record family history data in text. The authors sought to characterize family history information within clinical documents to assess the adequacy of existing models and create a more comprehensive model for its representation. Models were evaluated on 100 documents containing 238 sentences and 410 statements relevant to family history. Most statements were of family member plus disease or of disease only. Statement coverage was 91%, 77%, and 95% for HL7 Clinical Genomics Family History Model, HL7 Clinical Statement Model, and the newly created Merged Family History Model, respectively. Negation (18%) and inexact family member specification (9.5%) occurred commonly. Overall, both HL7 models could represent most family history statements in clinical reports; however, refinements are needed to represent the full breadth of family history data.

INTRODUCTION

Continued improvements in links between genotypic and phenotypic information and advancements in the knowledge of the molecular and genomic basis of disease present great opportunities for family history information to assist with prediction and prevention of disease.[1 2] Despite increased interest in family history and availability of the electronic health record (EHR), electronic tools for storing and making use of family history within many EHR systems are limited[3] and family history collection is often inadequate by clinicians.[4]

Several important standards, including HL7 version 3 Clinical Statement Model[5] and Clinical Genomics Family History Model,[6 7] provide a means for family history data representation. There is ongoing consideration for replacement of the HL7 Clinical Genomics Family History Model 'ClinicalObservation' with the HL7 Clinical Statement model,[6] and the synchronization of both HL7 models for family history representation will be increasingly important. Initiatives, perhaps most prominently, *My Family History Portrait*,[8] allow for web-based entry of structured family history by patients. In addition, the American Health Information Community Family Health History Workgroup recently outlined a core family history dataset for future EHR systems.[3]

Concurrent with this, large amounts of family history data continue to be recorded electronically in text format. The use of narrative or text occurs with documentation and communication throughout medicine with patients, family members, colleagues, and the scientific community, such as journal publications. Narrative within clinical documents helps clinicians synthesize complex facts and data elements and allows expression in a manner easily interpreted by other clinicians. In contrast, structured data can be difficult to interpret by clinicians due to loss of contextual information.[9] Structured data, however, is valuable for computer-based functions like decision support, administrative functions, and research.

CASE DESCRIPTION

The work described in this paper is part of a greater group effort aimed at the development of an effective natural language processing (NLP) tool to mine family history from clinical documents. Previous studies that used NLP for family history extraction from text[10 11] focused on family member and disease identification. We realized that clinical narrative of family history contains additional complex semantic elements. An important foundational step was to characterize family history data better within clinical reports and to evaluate the adequacy of existing models to aid in use of these data.

METHODS

Family history representation with HL7 models

We identified two standards with the capacity to represent family history information. First, the HL7 Clinical Genomics Family History Model was designed to assist as an intermediary representation of family history information from genetics applications (eg, Progeny), early-adopter EHR systems, and consumer empowered tools.[6 7] In addition to family history information, this representation facilitates diverse applications including pedigree construction, genomic information modeling, and disease risk assessment. Second, HL7 version 3 also includes the *Clinical Statement Model*[5] which seeks to harmonize clinical statement requirements into a single model for different specifications (such as the Clinical Document Architecture (CDA)[12 13]).

As such, a clinical document represented overall with HL7 CDA would have entries represented with the Clinical Statement Model. In contrast to the Clinical Genomics Family History Model, the Clinical Statement Model was designed to be a generic model to represent many different types of clinical statements and not specifically to represent family history information.

Both HL7 models were analyzed by three reviewers (two formally trained informaticians and one medical doctor with informatics training) for family history related data elements that could be represented. Model specifications were obtained from detailed analysis of the online HL7 standard documentation. Areas of inadequate coverage were identified from our analysis of the two models and analysis of an initial set of 50 clinical documents then combined into a Merged Family History Model for the final evaluation.

Clinical document analyses

This study utilized clinical documents from the University of Minnesota affiliated Fairview Health Services for 2002–08. Our corpus included all inpatient admission notes, inpatient consults, and outpatient consults because these documents are typically in the format of a history and physical examination with family history sections. University of Minnesota institutional review board approval was obtained and informed consent waived for this study.

Several preparatory steps were performed to obtain family history statements from clinical documents. All sentences with relevance to family history, including those not within the family history section, were analyzed. Sentences were divided into individual statements. Statements were defined as individual discrete items of clinical information from the sentence. Issues about inclusion or exclusion of sentences and individual statements were settled by consensus.

We utilized a primary test set of 50 random documents from the corpus to identify data content and structure of family history statements from clinical documents and as a tool to create a more comprehensive model for family history representation. Each statement was examined for data content and a representation of the structure of each statement was developed. We then used this generic structure from the 50 documents and the findings from our analysis of the HL7 Clinical Statement Model and HL7 Clinical Genomics Family History Model to create a Merged Family History Model.

Finally, we evaluated the adequacy of coverage of the HL7 Clinical Statement, HL7 Clinical Genomics Family History, and Merged Family History Models on 100 documents. Content of each family history statement was placed into a structured format by two reviewers (one health informatics graduate research assistant and one physician). If a data element did not suggest or fit into one of the defined fields, this was recorded as part of the error analysis. Each statement was then assessed for model coverage (Yes/No). Because the primary aim was to examine models for sufficient structure, it was assumed that incompletely specified family members (eg, 'uncle') and diseases (eg, 'colon problems') could be coded. Inter-rater reliability of the two reviewers for sentence inclusion and coding task was assessed on 20 documents with the κ statistic.

EXAMPLE

Family history representation with HL7 models

Table 1 summarizes our findings with respect to both HL7 models for representing family history. With the Clinical Genomics Family History Model, there are two types of statement structures if information is associated or not with a family member. This model also allows for statements describing family members without associated diseases and capacity to specify age ranges when the exact age is not known.

In contrast, the HL7 Clinical Statement Model also allows for statements of family history with or without specifying a family member but requires the statement to be associated with an Observation (ie, condition/disease). This model lacks specifications for current age, age of death, and age of diagnosis.

Neither specification appears to have explicit representations of 'unknown', 'unremarkable', or 'non-contributory' family history, the concept 'side of family', or the concept of 'adopted'. Neither model could represent concepts such as 'old age', 'elderly', and 'early age,' which were encountered in our initial set of cases. Ethnic origin of the patient or relative could be represented in both models, as could concepts of multiples or twins (not in table 1 or in the initial document corpus). Though 'healthy' is not a disease, the concept 'healthy' can be represented in both models as an Observation.

Defining family history data within clinical reports

In the 50 document test set, there were 131 sentences with family history data (15 (11%) located outside the family history section), resulting in 243 total statements (median 1, range 1–6 per sentence). A majority of statements fit into two data patterns (table 2), either family member and disease (family member disease statement) or disease with no family member mentioned (general family history disease statement). Some statements had information about family members (existence of, alive/deceased status, and findings of 'healthy') without reference to disease (family member statement) or were statements about family history without reference to disease (general family history statement). Table 2 also has examples of each pattern. Three statements were classified as 'other', as they did not fit into the four most common categories. Statements about a family member or a disease could include modifying information or further details such as family member status, current age, or age of disease diagnosis (table 3). Family members, diseases, and ages were sometimes inexactly specified. The 50 clinical documents and two HL7 model analyses were then used to create a Merged Family History Model.

Family history information representation coverage evaluation

A coverage evaluation for both HL7 models and the Merged Family History Model utilized 100 documents (52 inpatient setting) with distribution among internal or family medicine (n=31), surgical specialties (n=35), and medical specialties (n=34). In total, there were 238 sentences with family history data, resulting in 410 statements. Table 4 summarizes findings with respect to location in text, use of negation, specification of family members, and statement coverage for each model. The HL7 Clinical Statement Model had the lowest performance with respect to coverage (77%), followed by the HL7 Clinical Genomics Family History Model (91%) and the Merged Family History Model (95%). Unique gaps in the HL7 Clinical Statement Model that resulted in poorer performance compared to the other two models were its inability to represent age of diagnosis, current age, and age of death. Both HL7 models had coverage gaps for statements referring to side of family.

Inter-rater reliability between the two reviewers for the subset of 20 documents for inclusion of sentences yielded a κ of 0.97 (proportion agreement 99%). Analysis of the coding task for the

Table 1 Data elements in HL7 Clinical Genomics Family History Model and HL7 Clinical Statement Model

Data element	HL7 Clinical Genomics Family History Model	Family member	No family member*	HL7 Clinical Statement Model	
Family history ID	FamilyHistory.id	O	O		
Observed	FamilyHistory.effectiveTime, ClinicalObservation. effectiveTime	O	O	Observation.effective Time	O
Status	FamilyHistory.statusCode, ClinicalObservation.statusCode	O	O	Observation.statusCode	O
Method	FamilyHistory.methodCode, ClinicalObservation. methodCode	O	O	Observation.methodCode	O
Side of family		G	G		G
Family member					
Name, code, code system§	Relative.code	R	N/A	RelatedEntity.code‡	O
Gender	Person.Administrative GenderCode	O	N/A	Person.Administrative Gender	O
Race	Person.raceCode	O	N/A	Person.raceCode	O
Ethnicity	Person.ethnicGroupCode	O	N/A	Person.ethnicGroupCode	O
Date of birth	Person.birthTime	O	N/A	Person.birthTime	O
Observed age	DataEstimatedAge†	O	N/A		G
Living status	Person.deceasedInd	O	N/A	entryRelationship, Observation	O
Living	LivingEstimatedAge†	O	N/A		G
Deceased		O	N/A	entryRelationship, Observation	O
Age	DeceasedEstimatedAge†	O	N/A		G
Date	Person.deceasedTime	O	N/A		G
Condition					
Name, code, code system§	ClinicalObservation.code	O	R	Observation.code	R
Presence/absence	ClinicalObservation. negationInd	O	O	Observation.negationInd‡	O
Certainty	FamilyHistory.uncertaintyCode ClinicalObservation. uncertaintyCode	O	O	Observation.uncertainty Code	O
Diagnosis of disease					
Age	DataEstimatedAge†	O	N/A		G
Date	ClinicalObservation. effectiveTime	O	O	Observation.effective Time	O

*Statements without specification of a family member.
†Allows for a range of ages if an exact age is not known.
‡May be precoordinated in Observation.code.
§See HL7 Concept Descriptor (CD) data type for additional elements.
O, optional; R, required; G, gap in coverage; N/A, does not apply to data pattern.

HL7 Clinical Genomics Family History Model, HL7 Clinical Statement Model, and Merged Family History Model showed proportion agreement (96%, 97%, and 98%) and κ (0.75, 0.94, and 0.85), respectively.

Table 2 Family history statements in 50 clinical reports

	Sentences n (%)	Statements n (%)	Example statements
Family member disease statement	79 (60%)	153 (63%)	'His father died of small cell lung carcinoma' 'A sister was diagnosed age 45 with breast cancer'
General family history disease statement	33 (25%)	66 (27%)	'Family history is negative for hypertension' 'Strong history on paternal side for colon cancer'
Family member statement	11 (8%)	16 (7%)	'Unknown family history for grandparents' 'Son and daughter are in college ages 18 and 21' 'Her mother is alive and healthy'
General family history statement	5 (4%)	5 (2%)	'Unknown on father's side of family' 'Non-contributory' 'Reviewed and unremarkable'
Other statement	3 (2%)	3 (1%)	'Patient is adopted' 'The (family history) does not fit into a familial cancer syndrome pattern' 'Patient is of Korean origin'

DISCUSSION

The use of family history information for patient stratification of disease risk is an important and underutilized tool in medicine. Our analysis of family history information from clinical documents showed most family history data to be within the family history section, but other sections, particularly history of present illness and social history, can often contain some of this

Table 3 Structure for family history statements in 50 clinical reports

	Family member disease statement	General family history disease statement	Family member statement	General family history statement
Side of family (paternal/ maternal)	Optional	Optional	Optional	Optional
Uncertainty	Optional	Optional	Optional	Optional
Family member*	Required	—	Required	—
Current status (alive or dead)	Optional	—	Optional	—
Age of death*	Optional	—	Optional	—
Current age*	Optional	—	Optional	—
Disease*	Required	Required	—	—
Negation	Optional	Optional	—	—
Uncertainty	Optional	Optional	—	—
Age of diagnosis†	Optional	Optional	—	—

*Family member or its modifiers and disease can be incompletely specified.
†Can be a range.

Table 4 Family history data in 100 clinical reports

	N (%)
Total statements	410 (100)
Statements not in family history section	37 (9.0)
Statements with negation	72 (18)
Statements with inexact family member	39 (9.5)
Statement types	
Family member disease statement	246 (60)
General family history disease statement	92 (22)
Family member statement	47 (12)
General family history statement	11 (2.6)
Other	14 (3.4)
Model coverage of statements	
HL7 clinical statement model	317 (77)
HL7 clinical genomics family history model	374 (91)
Merged family history model	388 (95)

information. Family history statements tend to primarily be statements of family members and diseases or statements about presence or absence of disease within the family. A significant proportion of statements contain negation and incompletely specified family members.

Both the HL7 Clinical Genomics Family History and Clinical Statement Model need further refinement to represent the concept of paternal/maternal side of family history of disease, general statements of overall family history (unknown or noncontributory), and non-specific ages (eg, 'early age', 'elderly'). The HL7 Clinical Statement Model also could not represent ages for different events (current age, age of death, and age of diagnosis), ranges of ages, and statements about family members without an associated observation (ie, construct pedigree without associated disease/condition). The authors believe the concept of side of family to be of particular importance, as this is a common clinical statement in medicine for disease patterns in families. Moreover, because patients often have incomplete details with respect to age of diagnosis, death, and current age, the authors believe that having at least the flexibility of the Clinical Genomics Family History Model to represent age ranges is valuable.

With respect to incomplete family member specifications, we observed that 9.5% of statements did not directly identify a specific family member. Proposed HL7 RIM harmonization of relative codes moves away from family members that are incompletely specified. While ideally this information is collected with complete detail, we believe that there remains value in maintaining flexibility in family member coding, rather than completely losing this information. Additionally, we believe that there is value in adding more restrictions with respect to use of coding systems (eg, SNOMED CT for disease) to deal with these issues. It may also be helpful to have restrictions that address pre- and post-coordination of data elements (eg, with negation).

Both HL7 models include data elements that were not explicitly included in this analysis. Further work will involve identification of additional elements to the Merged Family History Model. For example, these models include a concept of 'Informant' (source of information from which family history was collected), which might be coded as 'Clinical Document'. We also encountered issues of uncertainty in our analysis. While sometimes certainty of assertions was at issue (eg, 'Sister died of probable liver cancer'), there were cases where the uncertainty was due to the information being unknown (eg, 'Paternal (family) history unknown'). While both HL7 models have capacity to represent uncertainty, there would be benefit to providing explicit guidelines for dealing with these semantic differences. With familial estrangement and adopted individuals being relatively common, we believe that it is important for future model versions to include both concepts of adoption and unknown family history, as lack of information concerning family member(s) is not the same as negative or positive statements.

Our analysis of family history information within clinical documents showed a range of complex and inexact features. While the HL7 Clinical Statement and Clinical Genomics Family History Models allow for representation of most data within clinical reports, further refinements are needed to represent the full breath of family history data in clinical documents.

Acknowledgements We would like to thank the University of Minnesota Institute for Health Informatics and Fairview Health Services for support of this project.

Competing interests None.

Ethics approval University of Minnesota institutional review board approval was obtained and informed consent waived for this study.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Guttmacher AE,** Collins FS, Carmona RH. The family history—more important than ever. *N Engl J Med* 2004;**351**:2333—6.
2. **Rich EC,** Burke W, Heaton CJ, *et al*. Reconsidering the family history in primary care. *J Gen Intern Med* 2004;**19**:273—80.
3. **Feero WG,** Bigley MB, Brinner KM. New standards and enhanced utility for family health history information in the electronic health record: an update from the American Health Information Community's Family Health History Multi-Stakeholder Workgroup. *J Am Med Inform Assoc* 2008;**15**:723—8.
4. **Suther S,** Goodson P. Barriers to the provision of genetic services by primary care physicians: a systematic review of the literature. *Genet Med* 2003;**5**:70—6.
5. HL7 clinical statement. http://www.hl7.org/special/Committees/clinicalstatements/ and http://www.hl7.org/v3ballot/html/domains/uvcs/uvcs.htm.
6. HL7 clinical genomics family history. http://www.hl7.org/Special/committees/clingenomics and http://www.hl7.org/v3ballot/html/domains/uvcg/editable/POCG_RM000040UV.htm.
7. **Shabo Shvo A,** Hughes KS. Family history information exchange services using HL7 clinical genomics standard specifications. *Int'l Journal on Semantic Web & Information Systems* 2005;**1**:42—65.
8. **United States Department of Health & Human Resources.** Surgeon general's family history initiative. http://www.hhs.gov/familyhistory/. January 12, 2009.
9. **Patel VL,** Arocha JF, Kushniruk AW. Patients' and physicians' understanding of health and biomedical concepts: relationship to the design of EMR systems. *J Biomed Inform* 2002;**35**:8—16.
10. **Goryachev S,** Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. *AMIA Annu Symp Proc* 2008:247—51.
11. **Friedlin J,** McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. *AMIA Annu Symp Proc* 2006:925.
12. **HL7 Version 3 Ballot Site - January.** Clinical document architecture. http://www.hl7.org/v3ballot/html/infrastructure/cda/cda.htm, 2009.
13. **Dolin RH,** Alschuler L, Boyer S, *et al*. HL7 clinical document architecture, release 2. *J Am Med Inform Assoc* 2006;**13**:30—9.