Representation of Information about Family Relatives as Structured Data in Electronic Health Records

L. Zhou[1,2,3]; Y. Lu[1]; C.J. Vitale[1]; P.L. Mar[1,2,3]; F. Chang1; N. Dhopeshwarkar[1]; R.A. Rocha[1,2,3]

[1]Clinical Informatics, Partners eCare, Partners HealthCare System, Boston, MA; [2]Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, MA; [3]Harvard Medical School, Boston, MA

Keywords

Terminology, SNOMED, HL7, family history, electronic health records, natural language processing

Summary

Background: The ability to manage and leverage family history information in the electronic health record (EHR) is crucial to delivering high-quality clinical care.
Objectives: We aimed to evaluate existing standards in representing relative information, examine this information documented in EHRs, and develop a natural language processing (NLP) application to extract relative information from free-text clinical documents.
Methods: We reviewed a random sample of 100 admission notes and 100 discharge summaries of 198 patients, and also reviewed the structured entries for these patients in an EHR system's family history module. We investigated the two standards used by Stage 2 of Meaningful Use (SNOMED CT and HL7 Family History Standard) and identified coverage gaps of each standard in coding relative information. Finally, we evaluated the performance of the MTERMS NLP system in identifying relative information from free-text documents.
Results: The structure and content of SNOMED CT and HL7 for representing relative information are different in several ways. Both terminologies have high coverage to represent local relative concepts built in an ambulatory EHR system, but gaps in key concept coverage were detected; coverage rates for relative information in free-text clinical documents were 95.2% and 98.6%, respectively. Compared to structured entries, richer family history information was only available in free-text documents. Using a comprehensive lexicon that included concepts and terms of relative information from different sources, we expanded the MTERMS NLP system to extract and encode relative information in clinical documents and achieved a corresponding precision of 100% and recall of 97.4%.
Conclusions: Comprehensive assessment and user guidance are critical to adopting standards into EHR systems in a meaningful way. A significant portion of patients' family history information is only documented in free-text clinical documents and NLP can be used to extract this information.

Correspondence to:
Li Zhou, MD, PhD
Clinical Informatics, Partners eCare, Partners HealthCare System
93 Worcester Street, 2nd floor
Wellesley, MA 02481
United States of America
E-mail: lzhou2@partners.org
Phone: (+1)781–4168489

Appl Clin Inform 2014; 5: 349–367
DOI: 10.4338/ACI-2013-10-RA-0080
received: October 12, 2013
accepted: February 11, 2014
published: April 9, 2014
Citation: Zhou L, Lu Y, Vitale CJ, Mar PL, Chang F, Dhopeshwarkar N, Rocha RA. Representation of information about family relatives as structured data in electronic health records. Appl Clin Inf 2014; 5: 349–367
http://dx.doi.org/10.4338/ACI-2013-10-RA-0080

1. Introduction

Family history is an important component of medical records for identifying patients at high risk for developing certain diseases [1, 2]. Systematically gathering detailed family history is critical to delivering personalized clinical care. Information about a patient's relatives is an essential element of family history; therefore, it is important to accurately represent and process this information in electronic health records (EHRs). Hereafter, we will use "relative information" to refer to familial relationships including family members and relatives.

The purpose of this study is threefold. First, we evaluated standard terminologies for representing and encoding relative information. Second, we examined patient relative information in an EHR, including both a structured family history module and free-text documents. Lastly, because a large amount of family history information was only recorded in free-text documents, we extended a natural language processing (NLP) system to identify, extract, and encode such information.

1.1 Background

Many diseases are the result of inherited conditions or the interactions of genetic, environmental, and behavioral factors [2]. Although genome technology and genetic testing have become more sophisticated, accessible and affordable, the family history remains a cost-effective and well-proven tool for supporting individualized disease prevention, diagnosis, and treatment. The family history has been shown to help predict the individual risk of a variety of diseases such as colorectal cancer [3], heart disease [4], breast cancer [5], and type 2 diabetes [6], among many others [7, 8]. Managing and leveraging family history information in EHRs is crucial to delivering high-quality clinical care.

1.2 Information Models and Standard Terminologies

Recently, researchers in biomedical informatics have made substantial efforts in developing methodologies and standards for representing family history as structured data in EHRs. One example of a standardized model for family history is the HL7 Version 3 Pedigree/Family History Model developed by the HL7 Clinical Genomics Work Group [9, 10]. This model allows for standardized representation of relatives of arbitrary distance from the patient, either directly (e.g., *grandmother* associated directly to the *patient*) or indirectly (e.g., *mother* associated to the *mother* of the *patient*). The HL7 Version 3 Continuity of Care Document (CCD) standard also includes a family history section that contains data defining the patient's genetic relatives in terms of possible or relevant health risk factors that may impact on the patient's healthcare risk profile [11].

In the United States, enacted under the Health Information Technology for Economic and Clinical Health (HITECH) Act, Meaningful Use is a set of standards defined by the Centers for Medicare & Medicaid Services (CMS) Incentive Programs that governs the use of EHRs and allows eligible providers and hospitals to earn incentive payments by meeting specific criteria [12]. One of the menu objectives for Stage 2 of Meaningful Use of certified EHR technology is the ability to record patient family history as structured data [13, 14]. More specifically, it is required that more than 20 percent of all unique patients admitted to eligible hospitals or seen by eligible professionals during the EHR reporting period have structured data records for one or more first-degree relatives. It is also required that a patient's family health history is captured using SNOMED CT® International Release July 2012 and US Extension to SNOMED CT® March 2012 Release [15], or using the HL7 Family History (Version 3) standard [10]. In order to promote EHR vendors' adoption of information standards and to ensure system interoperability, the adequacy and coverage of the recommended standards must be fully assessed. However, to our best knowledge, there are no published studies showing the coverage of these two standards for representing and encoding family relative information in EHRs. We therefore conducted a comprehensive study to analyze how well the two standards are able to represent a sample of family histories documented in an ambulatory EHR system. Assessment of other standard terminologies that are used for representing and encoding family relative information, such as LOINC [16] and the UMLS Metathesaurus [17], were out of scope for this study.

1.3 Family Histories in EHRs

EHR systems typically offer different ways for clinicians to document patients' family history information as structured or free-text data.

1.3.1 Structured Data

The preferred option for gathering structured family history information is a dedicated "Family History Module" within the EHR. This module enables clinical users to capture family history and support genetics-related decision making. Using this module, the patient's family medical history and relative information can be recorded, including attributes such as the familial relation to the patient, age of onset of the disease, living status, and age of death. The system can then stratify the patient into different risk groups (e.g., low, moderate, high) based on the information provided, and suggest diagnostic screening options to the clinician. Potential drawbacks to using a structured data module versus free-text notes or dictated narrative include decreased familiarity, lack of ease of use and freedom to express anything the clinician wishes, and loss of mechanisms that augment or enrich simple facts, such as qualifying severity or degree, conveying temporal relationships, indicating patterns of causality, providing rationale, proposing hypotheses, and suggesting alternatives [18].

Other methods for collecting structured family history include "Family History Forms" in a Personnel Health Record (PHR), which allow the patient to self-report family history information [19].

1.3.2 Free-text Data

Despite the availability of dedicated EHR modules, family history information is frequently collected using a traditional approach by direct questioning from clinicians. The information is then captured in free-text documents such as clinic visit notes, admission notes, and discharge summaries [20]. Examples include "Her mother died at 88 of a myocardial infarction", "Her 47-year-old son is healthy", and "Mom d 69 w/ DM & HTN". These free-text documents serve as important data sources for gathering and integrating family history information.

1.4 Natural Language Processing (NLP)

Free-text family history information must be converted and represented in a structured format in order to be used for subsequent automated processing. While several previous studies reported which terminologies were used to encode problems and diagnosis, very few studies described how to formalize and encode relative information. Friedlin and Clement [20] conducted a study in which NLP was used to locate the family history section within a hospital admission note. They analyzed the NLP performance in identifying 12 diseases and relative degree. This system achieved a positive predictive value of 0.96 and sensitivity of 0.93 for identifying relative degree, but it is unclear how specific relatives were identified and classified. Gorychev et al. [21] developed an NLP algorithm within the HITEx [22] to extract family histories from discharge summaries and outpatient clinic notes. Family members were identified using UMLS concepts under the "family group" (T099) semantic type. This system achieved a precision of 0.96 and a recall of 0.93 in family history diagnoses detection, and 0.92 and 0.92, respectively, in specific family member assignment. Lewis et al. [23] developed a set of dependency patterns and used the Stanford NLP Parser [24] to map specific family members to specific diseases. This approach was able to achieve a precision of 0.82 and recall of 0.52, but details were not provided regarding their relative lexicon and encoding strategies. ConText [25], an NLP algorithm that handles contextual information from free-text clinical documents, includes a set of terms representing "Experiencer" contextual information, but these terms are not coded. In this study, we created a comprehensive lexicon containing diverse relative concepts and their lexical variations, and refined an NLP system developed at Partners Healthcare System, called the Medical Text Extraction, Reasoning and Mapping System (MTERMS) [26], to process and encode family relative information found in free-text clinical documents.

2. Methods

Our methods involve three major steps, corresponding to the threefold purpose of this study, as mentioned in section 1. This study was approved by the Partners HealthCare Human Research Committee.

2.1 Investigate Information Standards and Related Work

As the first step, our focus was on the analysis of HL7 Family History (Pedigree) Standard and SNOMED CT, including not only blood relatives of a patient, but also other persons in family (e.g., adopted or foster child, spouse, and domestic partner). The HL7 Clinical Genomics Work Group has developed a set of concepts representing familial relations in support of the HL7 V3 Family History (Pedigree) standard. We included the relative codes from the HL7 RoleCode vocabulary [27]. Relevant SNOMED CT concepts (International Release July 2012) were retrieved from the social context axis, particularly under sub-tree (subClassOf) "person in the family". In addition, we examined relevant terms from the ConText algorithm that specifies "Experiencer" information. We identified coverage and gaps of each source for representing and encoding relative information in EHRs. We also included an analysis of 0–3 relative degrees.

The following definitions were adopted to classify relatives into different degrees [28]:

- Zeroth-degree relative: A family member with nominally 100 percent of their alleles identical by descent with a given individual in the family (e.g., monozygotic).
- First-degree relative: A family member with nominally 50 percent of their alleles identical by descent with a given individual in the family (e.g., parents, offspring, siblings).
- Second-degree relative: A family member with nominally 25 percent of their alleles identical by descent with a given individual in the family (e.g., grandparent, grandchild, uncle, aunt, nephew, niece, half-sibling, etc.).
- Third-degree relative: A family member with nominally 12.5 percent of their alleles identical by descent with a given individual in the family (e.g., cousins, great aunts, great uncles, etc.).

2.2 Review of Family History Information in EHRs

In this step, we reviewed relative concepts recorded in both structured and free-text format as well as those concepts used for genetics-related decision making.

2.2.1 Local Relative Concepts in the Structured Family History Module

We reviewed local concepts for encoding relative information in an ambulatory EHR system, called the Longitudinal Medical Record (LMR), developed by and used at Partners HealthCare System in Boston, Massachusetts. The family history module in the LMR is used to capture patient family history and support genetics-related decision making. Current algorithms for risk stratification focus on five diseases, including breast cancer, coronary artery diseases, colon cancer, diabetes mellitus type II, and osteoporosis. Clinical decision support (CDS) rules in the LMR can be triggered based on a patient's family history information. For example, IF a female patient has ≥ 1 second-degree paternal female relative with breast cancer at age ≤40 years AND the patient's last mammogram was ≥ 12 months prior, THEN the CDS will suggest that the patient should have a primary care physician visit within 3 months. We reviewed the set local concepts for representing relative information in the LMR, and mapped these concepts to SNOMED CT and HL7 RoleCodes to verify their coverage.

2.2.2 Relative Information in Free-text Documents

This study involved 15,006 patients who visited the Brigham and Woman's Hospital (a founding member of the Partners HealthCare System) between June and December 2012. The patients' free-text clinical documents, including admission notes and discharge summaries, were requested from the Partner's Research Patient Data Repository [29]. The June 2012 data were used to refine MTERMS [26] for processing family relative information. Based on our review, we estimated that about 10–20% of admission notes and discharge summaries contained a family history section.

We then randomly selected 100 admission notes and 100 discharge summaries from the July to December 2012 data, all of which included a family history section header. These 200 documents were obtained from 198 individual patients. A physician and a PharmD candidate conducted a manual review of these documents for three purposes:

1. examine the occurrence and diverse expressions of relative information,
2. assess the coverage of the standards for encoding relative information, and
3. create a "gold standard" for evaluating the performance of MTERMS.

We further reviewed these patients' structured family history information available in the Family History Module of the LMR, and analyzed differences in documenting family history information in structured and unstructured format.

2.3 Develop and Evaluate an NLP Module for Processing Free-text Relative Information

MTERMS was modified and extended to automatically extract and encode relative information from the free-text documents. MTERMS locates the family history section from clinical documents by recognizing section headers and phrases at the beginning of a paragraph or statement based on a set of 78 different expressions of family history section headers (e.g., "FHx", "Family History Review", "Family/Social History", etc.). These expressions were compiled after the manual review of clinical documents and structured note templates used by the LMR. Manual review of the June 2012 data helped build the relative information lexicon. In particular, we included diverse expressions (e.g., abbreviations and common typos) that were not included in standard terminologies. We evaluated the performance of MTERMS in identifying family members and relatives from free-text notes using the gold standard described above. Standard evaluation metrics [30] such as precision and recall were calculated.

3. Results

3.1 Information Standards and Relative Concepts

Although Stage 2 of Meaningful Use certification criteria have defined that either terminology standard can be used to capture patient's family health history, the structure and content of these two standards are different in several ways.

As a comprehensive reference medical terminology, SNOMED CT includes a wide range of concepts representing person in family, including relative, adopted person, person in family of subject, and extended family member. ▶Figure 1– left shows the hierarchy under the class *person in the family* of axis *social context*. Under the relative hierarchy, *relative* is classified into *immediate family member, non-immediate family member, blood relative, distant relative, aunt, cousin, grand-parent, grand-child*, and so on. Given SNOMED CT's multi-hierarchical structure, a concept can have multiple parents. For example, *natural mother* is a *first degree blood relative* and an *immediate family member*. In addition, contained is a set of concepts representing *family history with explicit context* under the *situation with explicit context* axis, including *family history unknown, no family history of clinical finding, family history of procedure, family history with explicit context pertaining to a specific relative*, and so on (▶Figure 1 – right).

In the HL7 V3 Family History (Pedigree) model, the relative class is defined as "*links two people as in a personal relationship...*" where the character of the relationship must be defined by a *Personal Relationship Role Type* code. Using the values of this domain, it is possible to designate the relation to the patient, or to the patient's family member. ▶Figure 2 – left shows the recursive association of *Person* and *Relative*, which enables a hierarchical representation of a pedigree. HL7 Role Code vocabulary includes a set of relative codes (concepts) representing the *Personal Relationship Role Type* and also defines the relationships between these concepts (e.g., *Child specifies Family Member* and generalizes *Daughter, Foster Child*, etc.). The Role Code vocabulary currently contains 97 relative concepts that span up to six hierarchical levels. ▶Figure 2 – right shows some examples of these concepts for family relatives.

Compared to the HL7 Family History standard, which is more focused on familial relations, SNOMED CT contains more descriptive terms intended for use in clinical documentation, such as *sick relative, diabetic relative, working father*, etc. SNOMED CT is more extensive and contains some concepts that are not available in the HL7 Role Code vocabulary, such as 4th degree relatives (e.g., *great-great grand parents*) and others indicating the genetic status of a particular person (e.g., *twins, sperm donors*, and *surrogates*). In addition, SNOMED CT also contains concepts specifying the contextual information of a clinical finding, particularly representing negations, such as *family history unknown*. However, SNOMED CT does not yet contain concepts for representing *maternal/paternal uncle, maternal/paternal aunt, maternal/paternal cousin*, among others, which are available in the HL7 Role Code vocabulary.

We conducted a crosswalk between relative concepts in SNOMED CT and HL7 Role Codes, and also terms available in ConText, along with other lexical variations found in free-text clinical documents (see examples in ▶Table 1). ConText includes some common family relative terms and their possessive and plural forms, but lacks other terms that are included in standard terminologies, and lexical variations found in clinical documents.

By compiling reference standards with common synonyms and misspellings, we created a comprehensive lexicon of family relatives containing 414 unique concepts and approximately 1,400 terms. This lexicon is used by MTERMS to identify a patient's relative information from free-text notes. Our lexicon is available upon request.

3.2 Standards' Coverage for Local Relative Concepts

The LMR currently includes 23 specific concepts for representing relative information, particularly for the purpose of supporting CDS. Among these concepts, 6 are first-degree relatives (*mother, father, brother, sister, son,* and *daughter*), 16 are second-degree relatives (*maternal half-brother, maternal half-sister, paternal half-brother, paternal half-sister, maternal aunt, maternal uncle, paternal aunt, paternal uncle, maternal grandmother, maternal grandfather, paternal grandmother, paternal grandfather, nephew, niece, grandson, and granddaughter*), and 1 is a third-degree relative (*cousin*). SNOMED CT does not include the first 8 second-degree relative concepts listed above and the HL7 Role Code vocabulary does not include the first 4.

3.3 Distribution of Different Relative Information in EHRs

3.3.1 Free-text Data

Among the 200 clinical documents (from 198 patients) that contained a family history section, 58 (29%) did not contain specific family history information (e.g., "unknown" or "non-contributory"), 30 (15%) contained family history information without specific relatives mentioned (e.g., "positive for DM"), and 112 (56%) contained specific relative information. Among those documents containing specific relative information, on average, 2.6 relatives were mentioned per document.

3.3.2 Structured Data

For the same 198 patients that had clinical documents with a family history section, only 34 (17.2%) also had relative information recorded using the LMR family history module; among which, 31(15.7%) patients had structured information, corresponding to 101 coded observations total. A total of 10 free-text records were found for the remaining 3 patients.

3.3.3 Comparisons between Free-text and Structured Data

Whereas the structured records contained general family medical history, the free-text documents generally contained only family history pertinent to a specific hospital visit and related to the present diagnoses. Relative information was usually well characterized in structured entries. In contrast, vague terms (e.g., "multiple family members") were found in free-text documents. Additionally, in free-text documents clinicians often used negation if a patient does not have a family history, or if the family history is unknown. Discrepancies between structured entries and free-text data were also identified. For example, a patient's structured data indicated that both the patient's mother and maternal grandmother had intracranial aneurysms, but examination of the free-text note indicated that the patient's mother had intracranial aneurysm and grandmother had an abdominal aortic aneurysm.

3.3.4 Analysis at family relative level

Our manual review of free-text documents confirmed that 78.4% of family relatives mentioned were first-degree relatives, 15.5%, 2.4%, and 0.3% were second-degree, third-degree, and zeroth-degree relatives, respectively. In the structured family history module, 85.6% of relatives identified were first-degree, while the rest were second-degree relatives. Details about the occurrences of diverse relative information in both sources are shown in ▶Table 2. Concepts that are not included in SNOMED CT or HL7 Role Code vocabulary are also indicated.

3.4 Coverage Rates of the Standards for Encoding Relative Information

For the 111 entries (34 patients) with structured family history records, SNOMED CT provided 95.5% coverage of relative terms, while HL7 Role Codes achieved 100% coverage. However, neither terminology standard was able to fully represent the relative information stored in free-text documents, with a coverage rate of 95.2% and 98.6%, respectively. Some of the missing terms include zeroth degree relatives (e.g., HL7 does not contain *identical twin*) and second-degree relatives (e.g., SNOMED does not contain *maternal/paternal aunt/uncle*). These concepts are vital to the development of robust CDS rules and must be included in standard terminologies. Additional details regarding coverage for encoding specific relatives can be found in ▶Table 3.

3.5 NLP Performance

Our evaluation of MTERMS based on the 291 relatives mentioned in the 112 documents demonstrated that it was able to successfully identify and encode relative information mentioned in free-text documents, with a precision of 100%, a recall of 97.4%, and an F-measure of 98.7% (▶Table 4). The five relatives missed were mainly due to the following reasons. Additional similar examples were provided to elucidate the challenges.

- *Misspelling or ill-formatting of free-text notes:* The function to detect word boundary should be improved to further enhance the performance of MTERMS. For example, "Father died of cancer; siblings and childrenhealthy" (missing a space between words).
- *Abbreviations/Acronyms:* Efficient algorithms should be developed to support word-sense disambiguation, helping to identify the right meaning of an ad hoc acronym. For example, GF may represent "grandfather" or "girlfriend".
- *Co-reference using person's name:* Correct relative recognition should include the function to recognize and assign names to specific relatives. For example, "His 2 daughters, Karen and Diane live within walking distance of his home". Later in the narrative, a clinician may repeat the names to refer to the patient's daughters.
- *Self-referencing terms:* some terms are used to describe the patient or the relative's genetic status, instead of interpersonal relationship, such as "he is the only child in the family".
- *Descriptive adjectives before noun relative terms:* we encountered terms such as "full siblings", "half siblings", "maternal great aunt", etc. These nuances should not be ignored because they indicate genetic difference. In addition, "younger", "youngest" and "another" were used in free-text notes to indicate that there are multiple entries for the same concept; however, it was difficult to identify and code this information.

4. Discussion

In this study, we reviewed information about family relatives from an ambulatory EHR, including structured and free-text formats. Our results showed that although the structured family history module was available in the EHR system, patient family history data were often not recorded in structured and coded format. Most family history information was only available in free-text clinical documents. We also found that family history information documented in different EHR locations can be inconsistent, confirming that clinicians will need to reconcile and update this information from time to time. With Stage 2 of Meaningful Use criteria, we expect that more family history information will become available in structured format. We also anticipate that specialized tools to help reconcile this information from different sources will become available, including pertinent information only found in free-text documents.

We reviewed the two standards required for the Stage 2 of Meaningful Use criteria to represent relative information in family histories. Although at current stage only first-degree relatives are required, it is necessary to conduct a comprehensive assessment of the standards to confirm their readiness to be widely adopted by EHR systems. Both SNOMED CT and HL7 RoleCodes contain gaps in representation of relative information stored in EHRs. For example, SNOMED CT lacks codes for *maternal uncle* and *aunt*, while HL7 lacks codes for *twins, donors*, and *surrogates*. Similarly, both reference standards lack codes for *maternal/paternal siblings*. Our research team had made a request to SNOMED CT to add the 8 relative concepts as mentioned in section 3.2. SNOMED CT accepted our suggestion to include these concepts in its upcoming version. As such, we predict an improvement in coverage rate for all free-text notes of 3.1% (from 95.2% to 98.3%).

When considering the representation of family relative information, SNOMED CT and HL7 Family History standard have a different purpose and scope. While both use a multi-hierarchical concept structure, the design of the HL7 relative RoleCodes is intended to enable a pedigree representation. Although SNOMED CT has a broader scope and includes more concepts about *person in the family* than HL7, our manual review of family history information found in free-text documents showed that SNOMED CT's coverage for representing relatives in family histories was lower than HL7's. However, SNOMED CT contains explicit concepts representing negative family history findings, while the HL7 relative RoleCodes does not. It is important to document negations as our analysis showed that a significant portion of patients' family history information was "unknown" or "no family history of a clinical finding."

Another important issue is the lack of detailed implementation guides that explain how the recommended standards should be used. For example, SNOMED CT and HL7 relative RoleCodes contain terms referring to 'generic' (primitive) and 'qualified' (pre-coordinated) relatives (e.g., "*mother*" vs. "*natural mother*"). Compared to "*natural mother*", "*mother*" is a generic concept that can refer to other family relative concepts, including "*natural mother*", "*adoptive mother*", "*legal mother*", "*step mother*" and "*surrogate mother*", and the meanings of these concepts are different. EHR systems may choose to adopt 'generic' and/or 'qualified' concepts, potentially compromising efficient data interpretation and system interoperability. Therefore, detailed specifications are needed to guide the meaningful application of interoperability standards.

Our comprehensive review of the standards to generate a more complete lexicon is a timely effort. It is not only critical for enriching the existing terminology standards, but also important for NLP systems designed to extract and encode family relative information found in free-text documents. Our review showed that family history information was mostly stored in free-text documents; therefore, using NLP to process into a structured and coded format will make these data available for subsequent use, particularly computerized decision support and research studies.

The performance of our NLP system, MTERMS, in identifying relatives from notes was satisfactory. However, to achieve higher recognition, future work is needed to resolve anaphora, deixis, and co-reference, as well as unspecific relative information that occurred commonly in free-text (e.g., maternal side).

This study has several limitations. First, the clinical documents that we reviewed were from Partners HealthCare System and may not represent the diversity and complexity of clinical documents. Second, the relative information that we studied was recorded using the family history module of LMR; relative information that appeared in other areas of the EHR was not included in this study. Third, we only reviewed admission notes and discharge summaries. Future work should include other types of notes, such as clinic visit notes and consultant notes. Finally, a single reviewer analyzed each free-text document and we did not measure inter-rater agreement. However, previous studies [21] reported that inter-rater agreement for annotating family member information in clinical notes was 100%; therefore, a single reviewer may be sufficient.

5. Conclusion

Our study showed the incompleteness of structured family history information in EHRs and the content gaps of existing standards for representing and encoding family relative information. Comprehensive evaluations and additional guidance is critical to ensure the adoption of standards into EHR systems in a meaningful and consistent way. Free-text documents represent as an important data source of patient family histories. The MTERMS NLP system was able to successfully extract and encode relative information from free-text clinical notes.

Conflicts of Interest

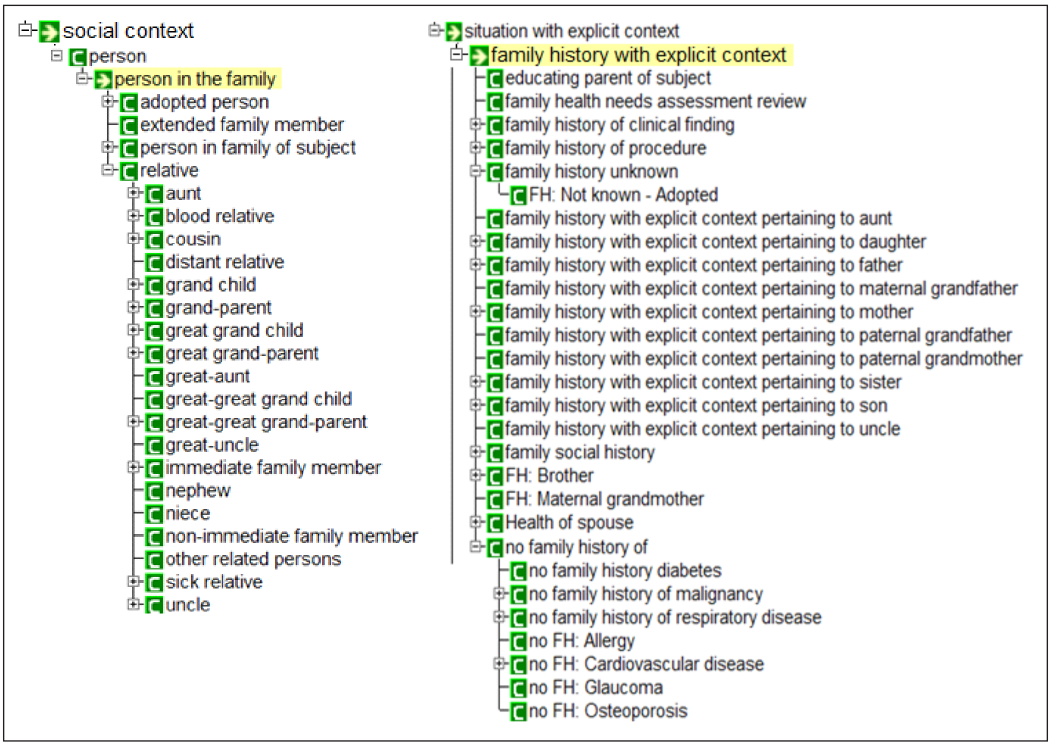
The authors declare to have no conflict of interest.

Protection of Human and Animal Subjects

This study was approved by the Partners HealthCare System Human Research Committee.

Acknowledgements

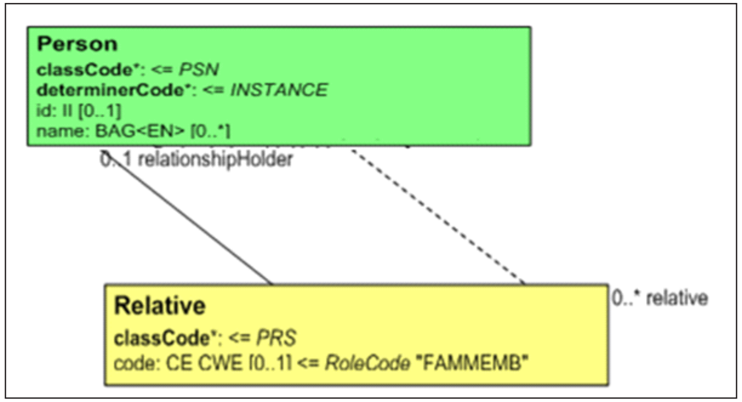
This project is supported by Partners-Siemens Research Council. The authors are grateful to Partners HealthCare System Quality Data Management team and Partners HealthCare System Research Patient Data Registry for their assistance in data retrieval.



- social context
  - person
    - person in the family
      - adopted person
      - extended family member
      - person in family of subject
      - relative
        - aunt
        - blood relative
        - cousin
        - distant relative
        - grand child
        - grand-parent
        - great grand child
        - great grand-parent
        - great-aunt
        - great-great grand child
        - great-great grand-parent
        - great-uncle
        - immediate family member
        - nephew
        - niece
        - non-immediate family member
        - other related persons
        - sick relative
        - uncle

- situation with explicit context
  - family history with explicit context
    - educating parent of subject
    - family health needs assessment review
    - family history of clinical finding
    - family history of procedure
    - family history unknown
      - FH: Not known - Adopted
    - family history with explicit context pertaining to aunt
    - family history with explicit context pertaining to daughter
    - family history with explicit context pertaining to father
    - family history with explicit context pertaining to maternal grandfather
    - family history with explicit context pertaining to mother
    - family history with explicit context pertaining to paternal grandfather
    - family history with explicit context pertaining to paternal grandmother
    - family history with explicit context pertaining to sister
    - family history with explicit context pertaining to son
    - family history with explicit context pertaining to uncle
    - family social history
    - FH: Brother
    - FH: Maternal grandmother
    - Health of spouse
    - no family history of
      - no family history diabetes
      - no family history of malignancy
      - no family history of respiratory disease
      - no FH: Allergy
      - no FH: Cardiovascular disease
      - no FH: Glaucoma
      - no FH: Osteoporosis

Fig. 1 Relative concepts in SNOMED CT: the "*person in the family*" hierarchy (left) and the "*family history with explicit context*" hierarchy (right)



Level	Code	Display
1	*(_PersonalRelationshipRoleType)*	
2	FAMMEMB	Family Member
3	PRN	Parent
4	FTH	Father
4	MTH	Mother
4	NPRN	natural parent
5	NFTH	natural father
6	NFTHF	natural father of fetus
5	NMTH	natural mother
3	CHILD	Child

Fig. 2 The association of Person and Relative in the HL7 Family History Model (left) and examples of HL7 RoleCode for family relatives (right)

Table 1 A crosswalk among relative concepts and negative family history in SNOMED CT and HL7 RoleCode vocabulary, terms in ConText, and other lexical variations found in free-text clinical documents

	SNOMED CT Concepts	HL7 Role Codes (value)	Con Text Terms	Synonyms, misspellings and other variations
Relatives by Degree				
0	identical twin sibling			identical-twin sibling
1	first degree blood relative			
1	sperm donor			
1	brother	BRO (brother)	brother('s), brother, brother's	bother, bro
1	older sister			older sis
2	grand-father	GRFTH (grandfather)	grandfather('s), grandfather, grandfather's	grand father, grandpa
2		MGRPRN (Maternal Grandparent)		maternal-grandparent
2	aunt	AUNT (aunt)	Aunt('s), aunt, aunt's	
3	female cousin			
Other Relatives				
	surrogate mother			surrogate mom
	adoptive daughter	DAUADOPT (adopted daughter)		adoptive dgtr
	legal parent			
Negation				
	Family history unknown	-	-	unknown
	No family history of	-	-	none

Table 2 Occurrences of relative information in free-text documents and structured FH module

	Free-text Documents			Structured FH Module			Concepts not in	
Concepts	Admission notes n (%)	Discharge summaries n (%)	Subtotal n (%)	Coded Entries n (%)	Free-Text Entries n (%)	Subtotal n (%)	SNOMED CT	HL7
Zeroth degree relative	**1 (0.8)**	**0**	**1 (0.3)**	**0**	**0**	**0**		
● Identical twin	1 (0.8)	0	1 (0.3)	0	0	0		*
1st degree relative	**97 (74.6)**	**131 (81.4)**	**228 (78.4)**	**89 (88.1)**	**6 (60.0)**	**95 (85.6)**		
● Parents	5 (3.8)	3 (1.9)	8 (2.7)	0	0	0		
● Father	27 (20.8)	35 (21.7)	62 (21.3)	28 (27.7)	2	30 (27)		
● Mother	27 (20.8)	32 (19.9)	59 (20.3)	34 (33.7)	2	36 (32.4)		
● Siblings	7 (4.6)	4 (2.5)	11 (3.8)	0	0	0		
● Sister	14 (10.8)	26 (16.1)	40 (13.7)	7 (6.9)	0	7 (6.3)		
● Brother	13 (10.0)	21 (13.0)	34 (11.7)	11 (10.9)	0	11 (9.9)		
● Children	1 (0.8)	2 (1.2)	3 (1.0)	0	0	0		
● Son	1 (0.8)	4 (2.9)	5 (1.7)	4 (4)	1	5 (4.5)		
● Daughter	2 (1.5)	4 (2.9)	6 (2.1)	5 (5)	1	6(5.4)		
2nd degree relatives	**23 (17.7)**	**22 (13.7)**	**45 (15.5)**	**12 (11.9)**	**4 (40.0)**	**16 (14.4)**		
● Maternal Grandparents	1 (0.8)	0	1 (0.3)	0	0	0		
● Grandfather	3 (2.3)	2 (1.2)	5 (1.7)	0	0	0		
● Grandmother	2 (1.5)	4 (2.9)	6 (2.1)	0	0	0		
● Maternal grandmother	2 (1.5)	2 (1.2)	4 (1.4)	4 (4)	1	5 (4.5)		
● Maternal grandfather	1 (0.8)	1 (0.6)	2 (0.7)	3 (3)	1	4 (3.6)		
● Paternal grandmother	1 (0.8)	2 (1.2)	3 (1.0)	0	1	1 (0.9)		
● Paternal grandfather	1 (0.8)	0	1 (0.3)	1 (1)	0	1 (0.9)		

Table 2 Continued

	Free-text Documents			Structured FH Module			Concepts not in	
Concepts	**Admission notes n (%)**	**Discharge summaries n (%)**	**Subtotal n (%)**	**Coded Entries n (%)**	**Free-Text Entries n (%)**	**Subtotal n (%)**	**SNOMED CT**	**HL7**
● Uncle	1 (0.8)	1 (0.6)	2 (0.7)	0	0	0		
● Paternal uncle	0	2 (1.2)	2 (0.7)	2 (2)	0	2 (1.8)	*	
● Maternal uncle	0	1 (0.6)	1 (0.3)	0	0	0	*	
● Aunt	2 (1.5)	3 (1.9)	5 (1.7)	0	0	0		
● Maternal aunt	4 (3.1)	2 (1.2)	6 (2.1)	1 (1)	1	2 (1.8)	*	
● Paternal aunt	1 (0.8)	0	1 (0.3)	1 (1)	0	1 (0.9)	*	
● Nephew	1 (0.8)	1 (0.6)	2 (0.7)	0	0	0		
● Grandchild	1 (0.8)	0	1 (0.3)	0	0	0		
● Granddaughter	1 (0.8)	0	1 (0.3)	0	0	0		
● Half-brother	1 (0.8)	1 (0.6)	2 (0.7)	0	0	0		
3rd degree relative	**4 (3.1)**	**3 (1.9)**	**7 (2.4)**	**0**	**0**	**0**		
● Cousin	2 (1.5)	2 (1.2)	4 (1.4)	0	0	0		
● Paternal cousin	1 (0.8)	0	1 (0.3)	0	0	0	*	
● Great aunt	1 (0.8)	0	1 (0.3)	0	0	0		
● Maternal great aunt	0	1 (0.6)	1 (0.3)	0	0	0	*	*
Others*	**5 (3.8)**	**5 (3.1)**	**10 (3.8)**	**0**	**0**	**0**		
● Wife	1 (0.8)	0	1 (0.3)	0	0	0		
● Girlfriend	1 (0.8)	0	1 (0.3)	0	0	0		
● Adopted	1 (0.8)	0	1 (0.3)	0	0	0		
● Mother's side	0	1 (0.6)	1 (0.3)	0	0	0	*	*

Table 2 Continued

	Free-text Documents			Structured FH Module			Concepts not in	
Concepts	**Admission notes n (%)**	**Discharge summaries n (%)**	**Subtotal n (%)**	**Coded Entries n (%)**	**Free-Text Entries n (%)**	**Subtotal n (%)**	**SNOMED CT**	**HL7**
● Father's side	0	1 (0.6)	1 (0.3)	0	0	0	*	*
● Family members	2 (1.5)	3 (1.9)	5 (1.7)	0	0	0		
Total	**130**	**161**	**291**	**101**	**10**	**111**		

*Others included blood relatives and non-blood relatives. Girlfriend was included as a domestic partner, which is defined as "the player of the role cohabits with the scoping person but is not the scoping person's spouse".

Table 3 Coverage rates of standards (SNOMED and HL7) for encoding relative information

	Admission notes			Discharge summaries			All documents			Family History Module		
	N	SNOMED CT n (%)	HL7 Role Codes n (%)	N	SNOMED CT n (%)	HL7 Role Codes n (%)	N	SNOMED CT n (%)	HL7 Role Codes n (%)	N	SNOMED CT n (%)	HL7 Role Codes n (%)
Zeroth degree relative	1	1 (100)	0 (0)[a]	0	-	-	1	1 (100)	0 (0)[a]	0	-	-
1st degree relative	97	97 (100)	97 (100)	131	131 (100)	131 (100)	228	228 (100)	228 (100)	95	95 (100)	95 (100)
2nd degree relative	23	18 (78.3)[b]	23 (100)	22	17 (77.3)[b]	22 (100)	45	35 (77.8)[b]	45 (100)	16	11 (68.8)b	16 (100)
3rd degree relative	4	3 (75.0)[c]	4 (100)	3	2 (66.7)[d]	2 (66.7)[d]	7	5 (71.4)[c,d]	6 (85.7)[d]	0	-	-
Others	5	5 (100)	5 (100)	5	3 (60.0)[e]	3 (60.0)[e]	10	8 (80.0)[e]	8 (80.0)[e]	0	-	-
Total	130	124 (95.4)	129 (99.2)	161	153 (95.0)	158 (98.1)	291	277 (95.2)	287 (98.6)	111	106 (95.5)	111 (100)

[a] HL7 lacks coding for twins; [b] SNOMED CT lacks coding for maternal uncle, maternal aunt, paternal uncle and paternal aunt; [c] SNOMED CT lacks coding for paternal cousin; [d] Both SNOMED CT and HL7 lack coding for maternal great aunt; [e] Both SNOMED CT and HL7 lack coding for maternal side and paternal side.

Table 4 MTERMS system performance on processing relative information (n = 291)

	Admission notes			Discharge summaries			All documents		
	Precision (%)	Recall (%)	F-measure (%)	Precision (%)	Recall (%)	F-measure (%)	Precision (%)	Recall (%)	F-measure (%)
Zeroth degree relative	100	100	100	100	100	100	100	100	100
1st degree relative	100	97.9	99.0	100	98.5	99.2	100	98.2	99.1
2nd degree relative	100	95.7	97.9	100	93.2	96.6	100	94.4	97.2
3rd degree relative	100	100	100	100	66.7	83.3	100	83.3	91.7
Others	100	100	100	100	100	100	100	100	100
Overall	100	97.7	98.9	100	97.2	98.6	100	97.4	98.7

Reference

1. Guttmacher AE, Collins FS, Carmona RH. The family history – more important than ever. New England Journal of Medicine 2004; 351(22): 2333–2336.
2. Feero WG, Guttmacher AE, Collins FS. Genomic Medicine – An Updated Primer. New England Journal of Medicine 2010; 362(21): 2001–2011.
3. Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE, Willett WC. A prospective study of family history and the risk of colorectal cancer. New England Journal of Medicine 1994; 331(25): 1669–1674.
4. Barrett-Connor E, Khaw K. Family history of heart attack as an independent predictor of death due to cardiovascular disease. Circulation 1984; 69(6): 1065–1069.
5. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA. Family history and the risk of breast cancer: a systematic review and meta analysis. International Journal of Cancer 1998; 71(5): 800–809.
6. Annis AM, Caulder MS, Cook ML, Duquette D. Family history, diabetes, and other demographic and risk factors among participants of the National Health and Nutrition Examination Survey 1999–2002. Preventing chronic disease 2005; 2(2): A19.
7. Scheuner MT, Wang SJ, Raffel LJ, Larabell SK, Rotter JI. Family history: a comprehensive genetic risk assessment method for the chronic conditions of adulthood. American journal of medical genetics 1997; 71(3): 315–324.
8. Valdez R, Yoon PW, Qureshi N, Green RF, Khoury MJ. Family history in public health practice: a genomic tool for disease prevention and health promotion. Annual review of public health 2010; 31: 69–87.
9. HL7 Clinical Genomics Work Group. The Family History Standard – Implementation Guide. November, 2012.
10. HL7/ANSI. HL7 Version 3 Standard: Clinical Genomics; Pedigree.
11. HL7 Implementation Guide: CDA Release 2 – Continuity of Care Document (CCD®). April, 2007.
12. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. New England Journal of Medicine 2010; 363(6): 501–504.
13. Centers for Medicare & Medicaid Services – EHR Incentive Program. Stage 2 Eligible Hospital and Critical Access Hospital Meaningful Use Menu Set Measures – Measure 4 of 6. Octobor, 2012.
14. Centers for Medicare & Medicaid Services – EHR Incentive Program. Stage 2 Eligible Professional Meaningful Use Menu Set Measures – Measure 4 of 6. October, 2012.
15. International Health Terminology Standard Development Organisation (IHTSDO). SNOMED Clinical Terms (SNOMED CT). 2012.
16. LOINC. http://loinc.org/ (Last accessed on 11/22/2013).
17. UMLS. http://www.nlm.nih.gov/research/umls/ (last accessed on 11/22/2013).
18. Johnson SB, Bakken S, Dine D, Hyun S, Mendonca E, Morrison F, et al. An electronic health record based on structured narrative. Journal of the American Medical Informatics Association: JAMIA 2008; 15(1): 54–64.
19. Surgen General's Family Health History Initiative: http://www.hhs.gov/familyhistory/index.html (last accessed on 10/6/2013).
20. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium 2006: 925.
21. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium 2008: 247–251.
22. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC medical informatics and decision making 2006; 6: 30.
23. Lewis N, Gruhl D, Yang H, editors. Dependency Parsing for Extracting Family History. Healthcare Informatics, Imaging and Systems Biology (HISB), 2011 First IEEE International Conference on; 2011: IEEE.
24. De Marneffe M-C, Manning CD, editors. The Stanford typed dependencies representation. Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation; 2008: Association for Computational Linguistics.
25. Chapman WW, Chu D, Dowling JN, editors. ConText: An algorithm for identifying contextual features from clinical text. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; 2007: Association for Computational Linguistics.
26. Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, et al. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium 2011; 2011: 1639–1648.

27. HL7 Version v3 Code System. http://hl7.org/fhir/v3/RoleCode (last accessed on July 16, 2013).
28. Genetics Home Reference. http://ghr.nlm.nih.gov/ (last accessed on 1/24/2014).
29. Partners' Research Patient Data Repository http://rc.partners.org/rpdr (last accessed on 10/6/2013).
30. Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. New York: ACM Press, Addison-Wesley. 1999.