Automated Extraction of Family History Information from Clinical Notes

Robert Bill[1], Serguei Pakhomov, PhD[1, 2], Elizabeth S. Chen, PhD[4, 5], Tamara J. Winden, MBA[1,6], Elizabeth W. Carter, MS[4], Genevieve B. Melton, MD, MA[1, 3]
[1]Institute for Health Informatics, [2]College of Pharmacy, and [3]Department of Surgery, University of Minnesota, Minneapolis, MN; [4]Center for Clinical & Translational Science and [5]Department of Medicine; University of Vermont, Burlington, VT; [6]Division of Applied Research, Allina Health, Minneapolis, MN

Abstract

Despite increased functionality for obtaining family history in a structured format within electronic health record systems, clinical notes often still contain this information. We developed and evaluated an Unstructured Information Management Application (UIMA)-based natural language processing (NLP) module for automated extraction of family history information with functionality for identifying statements, observations (e.g., disease or procedure), relative or side of family with attributes (i.e., vital status, age of diagnosis, certainty, and negation), and predication ("indicator phrases"), the latter of which was used to establish relationships between observations and family member. The family history NLP system demonstrated F-scores of 66.9, 92.4, 82.9, 57.3, 97.7, and 61.9 for detection of family history statements, family member identification, observation identification, negation identification, vital status, and overall extraction of the predications between family members and observations, respectively. While the system performed well for detection of family history statements and predication constituents, further work is needed to improve extraction of certainty and temporal modifications.

Introduction

Family history information is essential for understanding disease susceptibility and is critical for individualized disease prevention, diagnosis, and treatment (1,2). With greater and more widespread adoption of electronic health record (EHR) systems and mandates with Meaningful Use Stage 2 to utilize structured family history functionality (3), there is an increasing opportunity to perform secondary analysis of family history information. Important applications include researching genetic influences between family history and disease, performing association mining of the EHR, and supporting public health research (4). For instance, family history can be used to calculate odds ratios and relative risks using genotypic and environmental interactions (5) to estimate the level of risk for certain diseases, as well as signal the need for further consultation (e.g., genetic counseling) or diagnostic workup (e.g., preventative health screening).

We have previously analyzed the representation of family history information in the EHR (6) and characterized family history free text comments in an EHR family history module (7). Most recently, we have expanded upon this work and have used multiple sources to develop a more comprehensive family history representation model (8). While two previous studies have described preliminary approaches for automated family history extraction from clinical texts (9,10), these approaches did not include functionality to classify family history from other non-family history texts, nor were high-level linguistic features like predication or linguistic chunking used. Currently, most established clinical natural language processing (NLP) systems (e.g., MedLEE(11–13) and cTAKES(14)) are primarily focused on extracting named entities such as diseases, medications, or procedures or contextual information (15). Family history information is frequently expressed via relations between named entities such as family members or diseases and also may contain contextual information such as certainty and negation, as well as information about vital status and age modifiers. We sought to incorporate family history functionality into an open-source NLP system, BioMedICUS (16) and to evaluate the performance of this module for family history extraction.

Methods

Figure 1 depicts a broad overview of the NLP pipeline and the process for its evaluation. The system was developed as part of our larger BioMedICUS NLP system that uses the Unstructured Information Management Architecture (UIMA) framework. UIMA, originally developed by IBM, is now an open source Apache document annotation framework that provides functionality to write sets of analysis engines (pipelines) that annotate and process unstructured text. Authoring UIMA-based analysis engines involves design and development of algorithms that read document text and add annotations to it in progressive stages through a pipeline where each analysis engine has access to annotations added to the common analysis structure (CAS) by upstream analysis engines.

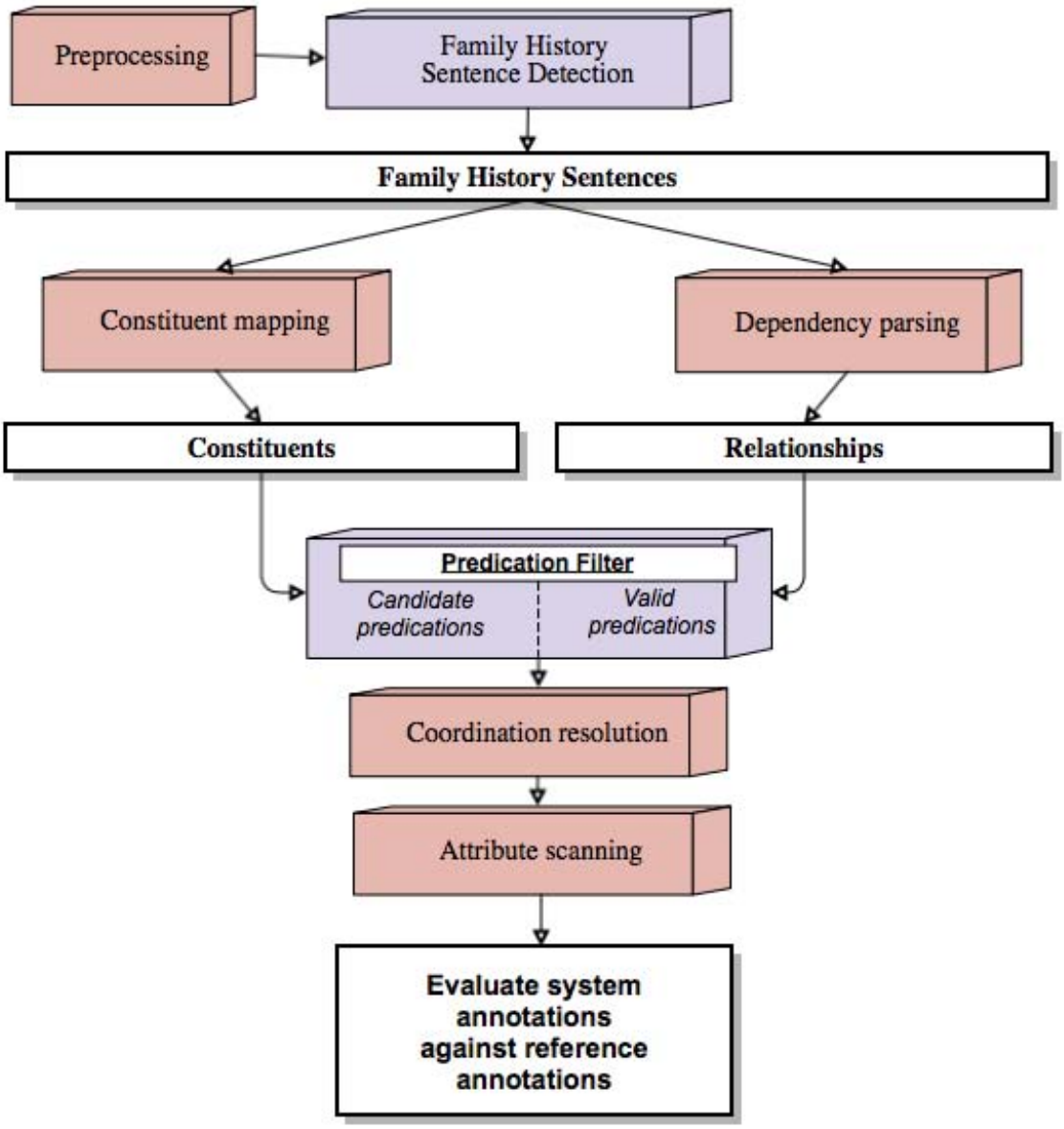

Figure 1. Overview of family history system and its evaluation

We used an approach to family history relationship extraction similar to the one developed by Rindflesch *et al.* for identification of semantic predications in biomedical literature (SemRep) (17). The NLP system shown in Figure 1 extracts family members and clinical observations as relationships where the relationship is usually identified by "indicator word(s)" (i.e., predication) appearing between the family member occurrence and clinical observation occurrence. After identifying the primary relationship, dependent phrases are then mined for attributes such as certainty, vital status and age. The pipeline was evaluated using an expert-based gold standard, as described below. Table 1 details each of the pipeline steps.

Corpora and Expert Annotation

The corpus used to develop and evaluate the family history NLP system was from a publicly available resource, MTSamples.com (18). For the MTSamples corpus, all History and Physical (H&P) notes, which were not behavioral health notes, were included (N=491). The family history pipeline was then developed using 329 MTSamples notes

(training set). As part of the evaluation, the model built for detecting family history statements was evaluated with 10 by 10 cross validation of the training set of notes. For this, the model was built with 90% of the notes and evaluated on the remaining 10%, with the remaining 10% being a different set of notes for each of the 10 iterations. Other portions of the pipeline using rules instead of models were evaluated on the entire training and test sets of notes. The system was then also evaluated on the remaining 162 MTSamples notes (evaluation set) as a more objective evaluation "hold out" set. The reference standard was developed by having annotators first identify family history statements in 20 notes to establish inter-annotator agreement prior to annotating the remaining set of MTSampels notes. After review of the general annotations completed in the first step, annotators added specific attributes. The specific attributes included family member, observation, side-of-family, vital status, certainty, and other temporal information. Table 2 shows an example set of annotations for the family history text for the reference standard.

Table 1: Family History Pipeline Components

Component	Description
Family History Sentence Detection	Determines if the current sentence contains a family history statement
Constituent mapping	Annotates the family history members from the HL7 clinical genomics family history model as family-member entities, and annotates all the disorders or procedures from SNOMED CT as observation entities
Relationship extraction	Identifies relationships between family member(s) and observation(s)
Coordination and list resolution	Transforms relationships that have lists of family members and/or observations into a list of relationships that specify a single family member and single observation. Also, adds attributes based on rules (e.g., adds "paternal-side" to relationships with "father" as the family entity)
Attribute scanning	Search patterns to identify the essential attributes of a relationship: vital status, certainty, age of diagnosis, side of family

Four informatics experts, including a physician, one PhD, one informatics graduate student, and one informatics researcher, provided manual annotations. Two of the experts also had previous experience in biomedical standard evaluation. Annotators arbitrated ambiguous sentences by discussing similar sentences to agree on presence of family history information. The resulting sets of annotations were provided in GATE (19) XML format and then converted to the UIMA XMI/CAS format for using in the UIMA pipeline developed for this study.

Table 2: Annotation Examples

Sentence 1: "He says his father might have died of heart disease."

Sentence 2: "Significant for epilepsy on the father's side of the family."

Sentence 3: "Mother with colon cancer but no other cancers."

	Family Member	**Observation**	**Vital status**	**Side of Family**	**Negation**	**Certainty**
Sentence 1	Father	heart disease	died			might
Sentence 2		epilepsy		paternal		
Sentence 3	Mother	colon cancer				
Sentence 3	Mother	other cancers			no	

Table 3 summarizes the annotations and the frequency of their occurrence in the corpus. If a family history statement included multiple clinical observations for a family member, then the UIMA annotators created all possible member-to-observation pairings. For example, the statement "*father died of stroke due to uncontrolled hypertension*" would result in the following pairs: *father-stroke* and *father-uncontrolled hypertension*.

The automated system was evaluated based on the accuracy with which it could extract these family member-observation pairings from text. Evaluation was challenged by a number of different complexities in the text, including the finding that many of the pairings were missing an explicit mention of a family member (e.g., "Significant for Alzheimer's.") The annotation for this statement does not include a family member; in this case, the automated system would then generate a pairing with a null value for the family member slot or position.

Table 3: Reference Corpus Annotations

Sentences	***Number***
Total Sentences	*23,155*
Family History Sentences	*284*
Predications	
Family member to observation predications	*364*
Attributes	
Side of Family	*15*
Family member	*417*
Observation	*745*
Vital status	*131*
Negation	*73*
Certainty	*55*

Family history sentence detection

Section and subsection headings contribute significant information when identifying statements of family history. For instance, short phrases such as "*Significant for asthma"* could be difficult to identify as family history statements when considering only the sentence. In these cases, the section heading, and sometimes the subsection heading provided the context necessary to improve accuracy in classifying family history statements. The current version of the BioMedICUS section detection module uses regular expressions to identify the section and subsection boundaries and headings. This is effective and efficient for corpora that have a limited set of consistent spacing and punctuation patterns that identify the sections and subsections such as the MTSamples corpus, although future versions could potentially benefit from augmented section heading tagging techniques. The heading label for sections and subsections are included with the sentence as predictors for family history sentence detection.

Using the section heading, subsection heading and sentence text as features together, sentences were classified as family history sentences. The system used a classification approach for family history statements detection using the SGD module in WEKA (20). WEKA is software for machine learning and predictive modeling, and the SGD module implements the stochastic gradient descent learning model for use in a support vector machine (SVM) (or

other linear models). N-gram-tokenized text is used as predictors to the SVM. This was compared to a simple lexical match that utilized all family words from the HL7 vocabulary RoleCode - PersonalRelationshipRoleType as part of the HL7 Clinical Genomics family history model (21). If a section or subsection heading contains the text "Family History," then all sentences within that section were then classified as family history statements. Only those sentences automatically determined to be family history statements were then evaluated further in the pipeline.

Constituent mapping and relationship mapping for core family history named entities and indicator words

The syntactic units that were candidates for the family history annotations were generally phrases found in either the HL7 Clinical Genomics family history model for family relation entities or in SNOMED CT as observation entities, which were extracted as described below. The family and observation entities were then linked together by an indicator word or phrase thus forming a predication relation. Indicator words are those words or phrases that signal or indicate a relationship between entities. They usually appear between the family entity and the observation entity; therefore, mapping indicator words was key to identifying family history statements. A lexicon of indicator words and phrases was constructed from the annotations provided in the corpus training set. Examples of these words are those that indicate possession or experience. The lexicon of possession words include *has, is, with,* and the tense variations. Experience indicators include *suffered a, died of, recovered from* and the tense variations of those words. Words and phrases in this indicator lexicon are also mapped during the mapping stage of the pipeline.

Together, the identification of the triples composed of the predication indicator, family member, and observation arguments, along with subsequent attribute extraction, formed the core functionality of the system, The SNOMED CT observation candidates were limited to only disorders and procedures using a disorder/procedure subset created by limiting the SNOMED CT concepts to those with semantic types defined by the UMLS Semantic Groups (22) as disorders and procedures. When searching for matches to family or observation in the corpus, only the longest matches (greedy matches) were kept. The mapping results for family member were also mapped to relative entities defined by HL7 family member codes.

Negation

For all sample statements that contained a negation term in the training corpus, negation occurred within a window to the left of an observation. Guided by this discovery, the negation detection implementation scans terms to the left of the constituent object in question. All negative results were then annotated. For example, phrases such as *no significant history* are annotated with the relationship between family and observation being noted as negated. Additionally, in the example, "*father, but not mother*", the negation component in the pipeline will add a negation attribute to only the *mother* entity of the phrase. The latter example also clarifies how scanning terms to the right of a constituent to determine their negation status can risk incorrect attachment of negations, such as the *father* constituent being identified as negated. Instead of attaching the negation to specific constituents of the sentence, the predication/relationship was annotated as having negation or no negation. Thus, the two sentences "*Neither parent has diabetes*" and "*Parents without diabetes*" would produce the same predications between parents and diabetes that have the same negation attribute.

Predication filter to convert family history constituents to valid relationships

As previously described, the family history annotation structure is composed of a relationship triple involving three entities: family member, observation and indicator word. The chunks on either side of the indicator word are assumed to be the concepts in the relationship. With initial exploration of the MTSamples training dataset, we observed using the family and observation syntactic units on either side of the indicator word or phrase was a reliable means of identifying entities defining a family history statement. Many of the family history statements in the training corpus include coordination in the family member and the observation. For example, the statement *Mother and father had diabetes and asthma* indicates four family history predications. The combinatorial considerations also include lists, such as *father, mother* and *uncle* or the disease list of *cancer, high blood pressure,* and *glaucoma*. These examples require coordination resolution to create all the necessary family history triples. The resolutions required in the previous examples are lexical. An additional issue that had to be addressed for family member chunks is that there were sometimes semantic lists created by statements such as *parents, three brothers,* and *all grandparents*. In order to manage these statements, a simple lexicon is used to lookup the appropriate combinations of family members for family member constituents that indicate multiple family members. The predication filter noted in Figure 1 removes all the candidate predications that are not valid because the dependency parse cannot confirm a relationship between the constituents and indicator.

At the end of this portion of the pipeline, a set of relationships, or triples, were generated and resolved -- each

having one family member, one observation, and one indicator per triple. The final modification to the relationship status is to evaluate the negation status of each entity. If any of the entities are negated, then the predication was negated. The training set did not contain instances of double negation; therefore, any negation in entities is sufficient to annotate the predication as negated. This defines only the relationship; therefore, subsequent steps in the pipeline extract attribute information from the indicator, and dependent phrases.

Attribute extraction and evaluation

Additional relationship attributes include vital status, certainty, age of diagnosis, and side of family. Some attributes are determined by the relationship itself, as in the example when the indicator chunk is an experiential frame such as *died of* (indicating the vital status attribute). Other attributes are found in dependent words or phrases. In order to accurately attach attribute words and phrases to the correct entity, the attribute extraction component used the Link Grammar dependency parser to identify dependent phrases. Each token in the dependent phrases is compared to a lexicon of terms that define attributes. Successful scans for certainty words, vital status words, and temporal references become attribute assignments. Figure 2 depicts an example sentence with the mapping from a sentence to chunks, followed by attribute assignment. Attribute results are reported as precision, recall, and F-score where the totals are cumulative across the corpus. The test corpus was evaluated as a whole as a hold out unseen set of notes.

Sentence: The patient states that both parents had heart disease and thinks that a maternal grandmother might have died of cancer.

Entities: familyEntity{both parents}
indicatorEntity{had},
observationEntity{heart disease}
familyEntity{maternal grandmother}
indicatorEntity{died of},
observationEntity{cancer}

Coordination: Rel1 = relationship{father, had, heart disease}
Rel2 = relationship{mother, had, heart disease}
Rel3 = relationship{grandmother, died of, cancer}

Attribution: sideOfFamily{Rel1, paternal}
sideOfFamily{Rel2, maternal}
vitalStatus{died of, Rel3}
certainty{might have, Rel3}

Figure 2. Relation and attribute identification

Results

Evaluation separated each of the stages of the pipeline. The first step of the pipeline, after general pre-processing such as tokenization and sentence segmentation, is the binary classification of sentences as a family history statement or not. This classification of family history sentences had good performance (Table 4), with the caveat that keyword detection was equally effective as the SVM approach, and that the performance depended upon the accuracy of sentence detection in general as well as accurate section heading detection. Sentences outside of the family history section were detected; however, section headings proved to be significant because statements within the family history section were frequently abbreviated with context derived only by the heading. For example, many family history statements included only the word "*noncontributory*," or observation lists such as "*diabetes, heart disease*." Section headings are the only means of classifying these abbreviated sentences, but family history statements in other sections do not appear in abbreviated forms. The results for accurately classifying sentences as family history statements include sentences within and without the family history section of the clinical note. A high number of false positive classifications of a sentence as a family history statement occur among sentences outside the family history section, particularly statements regarding family dynamics that include many family member terms.

Once sentences that contain family history statements are identified, the subsequent step is identifying the family history constituents, including the family member, observation and the indicator word used to identify predications. The different constituents of family history statements were evaluated using the set of sentences known to contain family history statements (Table 5). We found family member detection results of 90.8% precision and 94.0% recall (F-score 92.4%) for a sample of text containing family history statements and 177 family member names. Observation detection had 80.2% precision and 85.7% recall (F-score 82.9%) on the evaluation set containing 325 annotated clinical observations. Following the identification of constituents, the next step is extracting the relationship between constituents to form predications. Predication detection performance achieved an F-score of 65.1% with precision of 70.3% and recall of 60.6% in the training corpus, with lower results for the test corpus, as shown in Table 4. Negation performance showed 49.5% precision and 68.2% recall (F-score 57.3%) over a sample of family history statements that contained 37 negation annotations. The performance of vital status and age of death are also summarized in Table 5.

Table 4. Evaluation of family history NLP module

	Training Corpus Results (10 by 10 cross validation)			Evaluation Test Corpus Results		
	Precision	Recall	F-score	Precision	Recall	F-score
Sentence detection	84.5	60.0	70.2	83.2	55.8	66.9
Predication detection	70.3	60.6	65.1	66.3	58.1	61.9

Table 5. Evaluation of individual components on family history statements

	Precision	Recall	F-score
Family detection	90.8	94.0	92.4
Observation detection	81.1	85.7	83.3
Negation	49.5	68.2	57.3
Vital status	96.3	99.2	97.7
Age of death	94.0	87.5	90.6

Discussion

In this study, we constructed a family history UIMA-based NLP module, specifically an annotation engine that classifies sentences that contain family history statements and a pipeline to extract important information about family history from clinical texts. The novel portion of the study was the system's conversion of narrative text into a family history data type for each family history statement with the family-member plus observation in the clinical note as the core relationship and then leveraging the use of predication in our NLP module. We found that the initial classification of sentences depended significantly on section information and, as such, the success of classification was dependent on accurate section detection, sentence boundary detection, and constituent mapping (family members and clinical observations). Individual sentences about family history frequently were found to have insufficient information to classify them without considering the section headings. Statements such as "*noncontributory*", "*unknown*", and "*nonsignificant*" are poor predictors for family history classification by themselves. Sentence segmentation errors also presented challenges for accurate detection of family history statements because the erroneous segmentation can be key in separating family member and observation entities that otherwise would be part of the same family history relationship. Furthermore, incorrect clinical and family mappings contributed to errors in classifying family history sentences.

Overall, we found that the core functionality of extracting the triple of indicator word or phrase, family member entity, and observation entity had reasonable performance. We also identified mechanisms for future system improvement by differentiating patient history, social history and family history early in the pipeline. The family member and observation detection results lag the performance reported by Goryachev *et al.* (10) and Friedlin *et al.* (9) although these studies were more limited in their evaluation. For instance, the former achieved 93% sensitivity and 97% positive predictive value; however, comparability is limited because they used a broader classification of family member. The Friedlin study was a limited report (poster) describing a potential family history system with limited results. The work by Goryachev *et al.* is more directly comparable and achieved higher reported results with family member detection of 85.12% precision and 86.93% recall; observation detection of 96.30% precision and 92.86% recall; and correct family member assignment to diagnosis of 92.31% precision and 92.31% recall. The previous work limited 'diagnosis' to 8 UMLS semantic groups as opposed to the 19 UMLS semantic types encompassing the disorder and procedure semantic groups used in this study. The system also did not address other modifications such as vital status, age of death, and certainty.

The lower performance of family history sentence detection results with the test set compared to training set may indicate some over-fitting to the training data. An additional limitation of this study is that the effectiveness on different sets of notes is unknown, particularly if section heading detection is troublesome. The mapping process of observations was limited to SNOMED CT observations and procedures, and that appeared to be sufficient for the corpora used in the tests. The largest influence on classifying and mapping statements was the section detection. Section headings supply significant information in identifying family history statements, and an important step to our future work will be improving the section detection phase and concept mapping phases of this system. We also observed the following pattern of errors when classifying sentences: the longest and shortest sentences were misclassified most often. The reason for the misclassification of the longer sentences was due to the errors in sentence segmenting, and misclassification of the very short sentences was due most often due to exclusion of either the family member or the observation entity. We also observed a high rate of false negatives for predication detection attributable to errors in finding the arguments for the predication. Furthermore, in many cases the misclassification was due to presence of family related information in statements reporting social rather than family history (e.g., *Mother was a smoker*). Future work with social history extraction may be helpful in reducing these types of errors as well as extending family history functionality for extraction of additional attributes such as temporal elements to improve the extracted family history information at a more detailed level.

Conclusion

Detecting family history statements and mapping them to regularized annotation structures holds promise for information extraction from clinical notes for important downstream uses. The results of our pilot study show that extracting family history information from clinical notes is a complex and challenging task that lends itself to NLP approaches developed for relation extraction. Specific challenges consist of a number of hard NLP problems including resolution of coordination and co-reference that, in turn, rely on the accuracy of lower-level processes such as sentence and phrase segmentation and negation detection. In this pilot, we were able to achieve reasonable performance and to identify a number of areas for improvement. Our next steps will include further refinement of both lower-level and higher-level components.

Acknowledgements

The National Institutes of Health (1 R01 LM011364-01 NIH-NLM, 1 R01 GM102282-01A1 NIH-NIGMS, U54 RR026066-01A2 NIH-NCRR) and Clinical and Translational Science Award (8UL1TR000114-02) supported this work. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

References

1. Guttmacher AE, Collins FS, Carmona RH. The family history--more important than ever. N Engl J Med. 2004;351:2333–6.
2. Dick R, Detmer DE, Steen EB. The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition [Internet]. NAP. 1997. Available from: http://www.nap.edu/catalog.php?record_id=5306
3. Stage 2 - Centers for Medicare & Medicaid Services [Internet]. [cited 2014 Feb 20]. Available from: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Stage_2.html

4. Kukafka R, Ancker JS, Chan C, Chelico J, Khan S, Mortoti S, et al. Redesigning electronic health record systems to support public health. J Biomed Inform. 2007;40:398–409.
5. Ottman R. Gene-environment interaction: definitions and study designs. Prev Med (Baltim) [Internet]. 2010;25(6):764–70. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2823480&tool=pmcentrez&rendertype=abstract
6. Melton, Genevieve B and Raman, Nandhini and Chen, Elizabeth S and Sarkar, Indra Neil and Pakhomov, Serguei and Madoff RD. 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. J Am Med Informatics Assoc. 2010;17(3):337–40.
7. Chen ES, Melton GB, Burdick TE, Rosenau PT, Sarkar IN. Characterizing the use and contents of free-text family history comments in the Electronic Health Record. AMIA Annu Symp Proc [Internet]. 2012;2012:85–92. Available from: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=prem&NEWS=N&AN=23304276
8. Chen ES, Carter EW, Winden TJ, Sarkar IN MG. Development of a comprehensive family health history information model. AMIA Annu Symp. 2013;
9. 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.
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. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000;270–4.
12. Friedman C, Hripcsak G, Shagina L, Liu H. Representing Information in Patient Reports Using Natural Language Processing and the Extensible Markup Language. J Am Med Informatics Assoc. 1999;6:76–87.
13. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Informatics Assoc. 2004;11:392–402.
14. Savova, Guergana K and Masanz, James J and Ogren, Philip V and Zheng, Jiaping and Sohn, Sunghwan and Kipper-Schuler, Karin C and Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Informatics Assoc. 2010;17(5):507–13.
15. Chapman WW, Chu D, Dowling JN. ConText: an algorithm for identifying contextual features from clinical text. Proc Work BioNLP 2007 Biol Transl Clin Lang Process [Internet]. 2007;81–8. Available from: http://dl.acm.org/citation.cfm?id=1572392.1572408\npapers2://publication/uuid/48523D28-904A-48D8-AF74-BE9CFD2904F4
16. BioMedICUS [Internet]. Available from: https://bitbucket.org/nlpie/biomedicus
17. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003;36:462–77.
18. Transcribed Medical Transcription Sample Reports and Examples - MTSamples [Internet]. [cited 2014 Mar 4]. Available from: http://www.mtsamples.com/
19. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol. 2013;9.
20. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. ACM SIGKDD Explor Newsl [Internet]. 2009;11:10–8. Available from: http://dl.acm.org/citation.cfm?id=1656278
21. Shabo, Dr. Amnon D, Hughes, Kevis S. D. HL7 Clinical Genomics Family History Model Abridged. 2007.
22. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform. 2001;84:216–20.