Systematic Analysis of Free-Text Family History in Electronic Health Record

Yanshan Wang, PhD, Liwei Wang, MD, PhD, Majid Rastegar-Mojarad, Sijia Liu, Feichen Shen, PhD, Hongfang Liu, PhD
Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Abstract

Family history is an important component in modern clinical care especially in the era of precision medicine. Family history information in the Electronic Health Record (EHR) system is usually stored in structured format as well as in free-text format. In this study, we systematically analyzed a family history text corpus from 3 million clinical notes for the patients receiving their primary care at Mayo Clinic. Family members, medical problems, and their associations were analyzed and reported. Our findings showed a great agreement between positive/negated medical problems mentioned in the diagnosis report and the family history, as measured by observed agreement and random agreement. We also found that the family history of some medical problems existed up to 10~15 years prior to the diagnosis date of such problems. Finally two patient cases were studied to show the medical problems in the diagnosis and family history associated with the timeline.

Introduction

A large number of common diseases (e.g., cardiovascular diseases, cancers, and Alzheimer's disease) have been shown to be familial [1–6]. Family history information, which can reflect genetic susceptibilities of familial conditions, is essential in clinical care in the era of precision medicine. Additionally, family history captures shared environment and behavioral risk factors among the family members which are also important in clinical care [7–10]. Family history information has been utilized for risk assessment and stratification, clinical decision support, and clinical research [11,12].

Family history can be stored as structured and unstructured data in the Electronic Health Record (EHR) system. How to capture comprehensive family history information remains a research question [13]. We hypothesize that systematically analyzing family history information in clinical notes can provide more insight towards the underlying utility of family history in clinical care as clinicians tend to document information that assists their decision-making process (i.e., family history information recorded by clinicians may be more valuable for patient care than the family history information stored systematically).

In this paper, we provide a systematic analysis of the family history section in clinical notes leveraging natural language processing (NLP). We utilized all clinical notes for a cohort of patients who received their primary care at Mayo Clinic in Year 2013 to perform the analysis. Family members, medical problem mentions and their associations with the corresponding family history were analyzed and reported. In addition, we used observed agreement and random agreement to measure the agreement between positive/negated medical problem mentions in the diagnosis report and those in the family history section. Then we studied the prediction power of family history for diagnosis by analyzing the timeline of medical problems in family history and diagnosis.

Background and Related Work

Many studies focus on how to accurately extract information from the free-text family history using Natural Language Processing (NLP) [14–18]. Fridlin et al. developed a rule-based NLP system for extracting and coding family history data from hospital admission notes [16]. Goryachev et al. developed a simple rule-based NLP algorithm to identify and extract family history from discharge summaries and outpatient clinical notes [17]. Bill et al. proposed an Unstructured Information Management Application (UIMA)-based NLP module for automated extraction of family history [14].

Apart from extracting information from the free-text family history, a couple of studies have performed systematic analysis using structured family history information. For example, Marder et al. used the structured family history information for Parkinson's disease [19]; Silverman et al. investigated the application of the structured family history on genetic studies of Alzheimer's disease [20]; Endevelt et al. summarized the information in the structured family history [8]. However, few studies focus on systematically analyzing the information contained in the free-text family history. The study conducted by Chen et al. reported an analysis of free-text family history comments in the EHR [21]. Yet, there are two shortcomings in their study. First, the free-text family history data are the auxiliary comments on the structured family history. Those data supplement the structured data and may not contain complete family history information. Second, only cancer-related information was extracted and studied in their work.

Materials and Methods

Dataset

As we aim to systematically assess the associations between medical problems and family history information, the cohort we select includes patients who received their primary care at Mayo Clinic in Year 2013. We envision that their specialty care would also be provided by Mayo Clinic. The resultant corpus used in our study contains 3,224,427 clinical notes for a total of 115,710 patients with an average of 27.8 notes per patient. The dataset cannot be made public because it is private and contains protected health information (PHI).

Methods

In the Mayo Clinic EHR system, the free-text family history is recorded and stored in the "family history section" of each clinical note. The "family history section" can be identified by matching section header "Family History:20109" in the clinical notes where "20109" is the section ID. An NLP tool MedTagger [22], incorporated with previously developed family history identification method [18], is utilized in this study to extract two major kinds of information from the family history section: (1) assembled medical problems, and (2) family members. The family members are extracted according to the list in Table 1 by using regular expression patterns. We omit "paternal" and "maternal" modifiers for "grandfather", "grandmother", "aunt", and "uncle" in this study. Since the accuracy of NLP tools has been verified in the previous studies [9,18], it is excluded in this study.

Table 1. A list of family members.

Aunt	Brother	Child	Cousin	Daughter
Father	Grandfather	Grandmother	Grandparent	Mother
Parent	Sibling	Sister	Son	Uncle

Hierarchical clustering is implemented to analyze the association between medical problems and family members in the family history [23]. In this study, our analysis is based on the sentence-level co-occurrence information of medical problems and family members. We calculate the frequency of co-occurrence of each medical problem and each family member. Suppose the frequency of the *i*th medical problem for the *j*th family member is f_{ij}, the frequency vector of the *i*th medical problem for *n* family members can be written as $f_i = (f_{i1}, f_{i2}, , , f_{in})$. Then we apply agglomerative clustering for f_i through f_m where m is the number of medical problems. In the agglomerative clustering, two closest frequency data points are merged into one cluster according to the Euclidean distance defined as below:

$$d(f_i, f_j) = \sqrt{\sum_k (f_{ik} - f_{jk})^2}. \tag{1}$$

Subsequently, average linkage clustering is used to merge pairs of clusters according to the average distance between clusters. The average distance between cluster C_1 and C_2 is computed as follows:

$$d(C_1, C_2) = \frac{1}{\|C_1\|\|C_2\|} \sum_{f_i \in C_1} \sum_{f_j \in C_2} d(f_i, f_j). \tag{2}$$

By iteratively doing so, all the data points can be merged into a single cluster. The same clustering methods are applied to the *n* frequency vectors of family members for *m* medical problems. Finally, the association between medical problems and family members can be observed through the cluster hierarchy.

To analyze the association between medical problems in family history and those in diagnosis, we use the diagnosis sections of those patients who have family history. The same NLP algorithm is applied to extract medical problem mentions from the diagnosis sections. Observed agreement and random agreement are used to measure the agreements between the positive/negated medical problems mentions in diagnosis and those in family history. The observed agreement and random agreement measures are defined by:

$$observed\ agreement = \frac{a+d}{a+b+c+d}, \tag{3}$$

$$random\ agreement = \frac{m_a + m_b}{a+b+c+d}, m_a = \frac{(a+b)\times(a+c)}{a+b+c+d}, m_b = \frac{(c+d)\times(b+d)}{a+b+c+d}, \tag{4}$$

where a denotes the frequency of positive medical problems mentioned in family history while positive in diagnosis, b the frequency of positive medical problems mentioned in family history while negated in diagnosis, c the frequency of negated medical problems mentioned in family history while positive in diagnosis, and d the frequency of negated medical problems mentioned in family history while negated in diagnosis.

Results

Extraction of Family History

Out of 115,710 patients, 77,810 (67.2%) have family history sections in their EHRs. The ratio of patients with documentation of family history is much higher compared to 12% in Endevelt et al.'s study [8]. It implies that physicians at Mayo Clinic pay a lot of attention to the family history information. Those patients form a Family History (FH) Cohort. We retrieved the EHRs of FH Cohort from the Mayo Clinic data repository and extract the family history sections. This resulted in a corpus of 278,918 family history documents, which is denoted as FH Corpus hereafter. The family history is generally written as semi-structured texts, short sentences and narratives. Table 2 lists a few examples of the family history from the FH corpus.

Table 2. Examples of family history from the FH Corpus.

Prostate cancer - three Brothers Liver Disease--- Brother Pancreatic Cancer--2 Brothers HTN-- Father Elevated Chol/Trigs.---Brother, Sister MI, Stroke-- Father age 54
FATHER Father alive, 74, High blood pressure MOTHER Mother died at 74 from COPD (smoker), CHF, overweight SISTERS 5 sisters alive SONS 2 sons alive GRANDPARENTS Maternal Grandmother, deceased, *** years - died of "old age" Maternal Grandfather, deceased, unknown Paternal Grandmother, deceased, 70's, COPD (smoker) Paternal Grandfather, deceased, 70's, colon cancer
This is *** third pregnancy. Her first two pregnancies, through her previous husband, resulted in full-term females who are currently healthy at the ages of *** and ***. *** has a healthy ***-year-old brother whose partner is currently pregnant. She has a healthy ***-year-old brother who has no children. She has a healthy ***-year-old sister who has a healthy ***-month-old daughter. ***s father is healthy at the age of ***. Her mother suffers from type II diabetes at the age of ***. ***s mother has six siblings of whom one, a sister, has at least seven healthy children and has had three miscarriages, for which no reason was given. ***s partner, ***, is reportedly healthy at the age of ***. He has a healthy ***-year-old brother who has a healthy *** son. His mother is healthy at the age of ***. His father, age ***, reportedly has an adult-onset arrhythmia. There were no reports of mental retardation, learning disabilities, or birth defects. No family members had babies that were still born or died early. There were no reports of cancer before the age of ***. The remainder of the family history was non-contributory to today's discussion. There is no consanguinity reported between these families.

*** indicates de-identified information.

Gender and Age Statistics of the FH Cohort

Since the family history information recorded for a male is distinct from a female; and it also varies for patients at different ages, the gender and age statistics reported in this section can help understand the study cohort and corpus. Our first observation is that the family history appears significantly more frequent for females (58.5%) than for males (41.5%). This result is consistent with Endevelt et al.'s results [8]. Female patients, according to a study, tend to "have longer visits, ask more questions, get more information, receive more counseling, send and receive more emotionally-concerned statements and appear more involved in the interaction than male patients" [24].

Figure 1 demonstrates the distribution of age in the cohort. About 33.3% of patients are at age <30. Apart from age <30, age 40~49 (15.7%) and 50~59 (15.5%) are two age ranges with more patients than other ranges.

Family Members Mentioned in the Family History Section

Figure 2 shows the distribution of family members mentioned in the family history. Obviously, the number of "father" (24.7%) and "mother" (23.7%) is significantly larger than that of other family members since parental history of disease is highly associated with patient's health. Together with "father" and "mother", "grandfather" (7.6%), "grandmother" (10.0%), "sister" (8.9%) and "brother" (8.0%) contribute a total of 82.9% of family members. These family members are of great importance for understanding family health genealogy.

Previous study finds that medical, developmental and pregnancy outcomes of first-, second-, and third-degree relatives are the most useful family history for a patient [25]. First-degree relatives including parents, offspring, and siblings, have 50% shared genes with the patient. Second-degree relatives including aunts and uncles, grandparents, half siblings, nieces and nephews, inherit 25% genes identical to the patient. Third-degree relatives including cousins and great-grandparents, share only 12.5% of genes with the patient [26]. Therefore, physicians would consider more information of the first-degree relatives in the family history. Figure 2 validates this result by showing that the number of first-degree relatives accounts for 71.2%, the number of second-degree relatives 27.8%, and the number of third-degree only 1.0%.

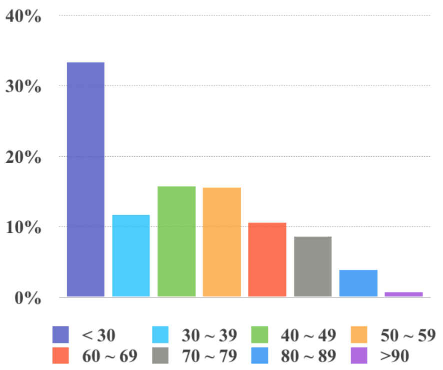


Figure 1. Distribution of age in the FH Cohort.

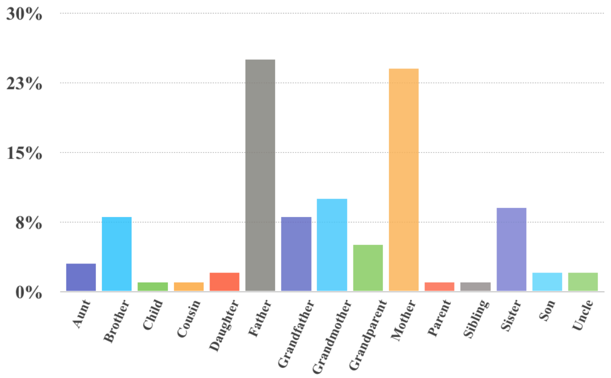


Figure 2. Distribution of family members in the FH Corpus.

Medical Problems in the Family History Section

There are 901,649 medical problem mentions corresponding to 9,646 unique medical problems extracted. The top ten frequent medical problems are listed in Table 3. Among the most frequent medical problems, hypertension, high blood pressure, high cholesterol and CAD are common cardiovascular diseases the family history has verified as a risk factor [10]. Diabetes and heart disease are early cardiovascular-related events [27]. Cancers are usually phenotypic diseases that can be revealed by the family history [28,29]. Thus, physicians are interested in those medical problems in the family. In addition, we find that physicians also pay attention to mental disorders in the family history (depression and alcohol abuse). This result is consistent with the result of some studies that family history may be enough to predict mental disorders due to the shared environment [30,31].

Family history may also improve the chances for early detection of rare diseases, since many rare diseases are gene-related medical problems. For example, hemophilia is an X-linked disease [32]. Apart from the frequent medical problems, a lot of rare diseases are also found in the free-text family history. For example, the frequencies of hemophilia and sickle cell anemia in the FH Corpus are 124 and 39, respectively. Though many standards and tools have been developed to gather information for common diseases in family history, there are currently no guidelines or standards on the collection of rare diseases at the point of care. However, the findings show that physicians at Mayo Clinic have paid attention to the family history of rare diseases.

Table 3. Top 10 medical problems in the family history section.

Mecial Problem	Frequency
Hypertension	49,460 (5.5%)
Depression	44,900 (5.0%)
Cancer	44,306 (4.9%)
Diabetes	35,281 (3.9%)
High Blood Pressure	33,437 (3.7%)
Alcohol Abuse	28,580 (3.2%)
Heart Disease	27,758 (3.1%)
High Cholesterol	25,592 (2.8%)
CAD	24,837 (2.7%)
Breast Cancer	23,266 (2.6%)

Association between Family Members and Medical Problems in Family History

Given the identified family members and medical problems in the FH Corpus (positive and negated), we would like to study the association between them, i.e., what medical problems are mostly considered for a specific family member. Using the hierarchical clustering for the frequencies of co-occurrence of medical problem and family member, we plot a heat map along with clusters in Figure 3.

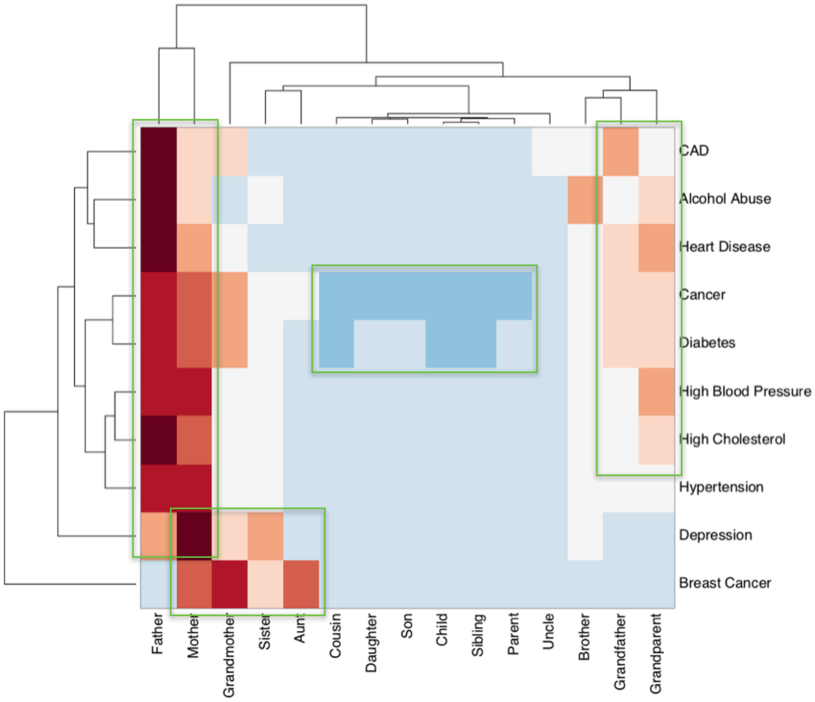


Figure 3. Heat map of frequencies of co-occurrence of top 10 medical problems and family members.

We have the following observations: (1) There are roughly four clusters indicated by green rectangles in Figure 3. (2) Almost all medical problems are considered for "father" and "mother". (3) Breast cancer frequently appeared for female family members ("mother", "grandmother", "sister", and "aunt") while alcohol abuse for both male and female family members ("father", "mother", "sister", "grandfather", "brother"). (4) Cancer and diabetes are clustered while high blood pressure, high cholesterol and hypertension are clustered. This is consistent with the known fact that cancer and diabetes are comorbidities [33,34]. (5) It is interesting that CAD, alcohol abuse, heart disease and high cholesterol are the most considered problems for "father" while depression is relatively the most considered for "mother".

In order to illustratively show the most concerned medical problems in genealogy, Figure 4 demonstrates a family tree where each family member is associated with the top 5 medical problems for that family member. Hypertension and depression are two mostly considered problems for each family member. CAD, diabetes, and MI are common mentioned problems for patient's siblings, parents and grandparents. Asthma is not among the top 10 medical problems in the FH Corpus but it is one of the most frequent medical problems for patient's siblings and children. So is colon cancer for grandfather, uncle and aunt; osteoporosis for mother; and ovarian cancer for aunt. Interestingly, alcohol abuse and mental illness are among top 5 considered medical problems for patient's son and daughter. This may due to the fact that people and their children usually live in a common environment, which is a key factor to both alcohol abuse and mental illness.

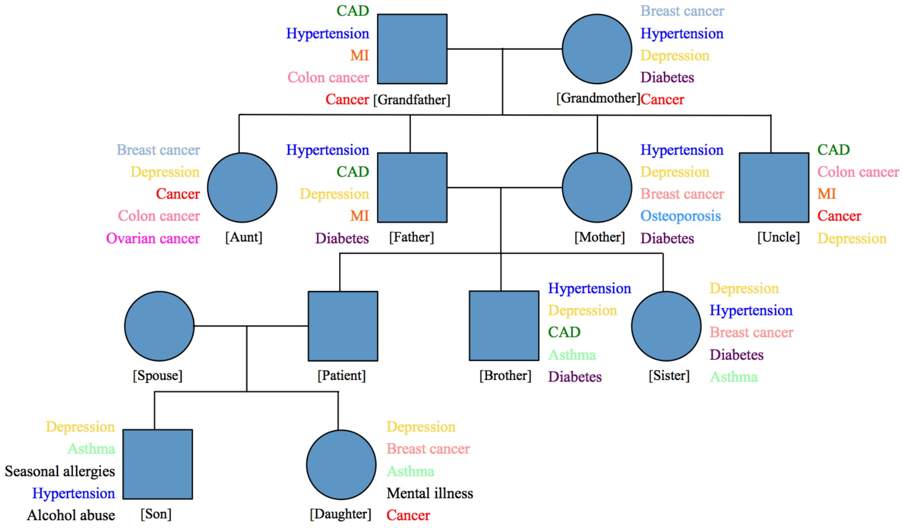


Figure 4. Top 5 medical problems for each family member in the FH Corpus.

Association between Medical Problems in Diagnosis and Those in Family History

In this section, we study the association between the medical problems in family history and those in patient's diagnosis reports. For each patient in the FH Cohort, we count the positive and negated medical problems mentioned in diagnosis reports. For each positive and negated mentions, we then check whether it is positive or negated in the family history and count the frequency. The accumulated results for the FH Cohort are summarized in a contingency table, as shown in Table 4. Observed agreement and random agreement are used to evaluate the agreement of positive and negated medical problems mentioned in diagnosis and family history. According to the definitions in Equations (3) and (4), the observed agreement and random agreement are 0.8100 and 0.7929, respectively. These measures indicate a great agreement between medical problems mentioned in diagnosis and family history. This result implies that the family history might have prediction power for the diagnosis.

We list the twenty most frequent positive medical problems in diagnosis while negated in family history and twenty most frequent negated medical problems in diagnosis while positive in family history in Table 5. For the "not found" medical problems, it is interesting that 95.9% of positive medical problems mentioned in diagnosis are not found in family history. Those "not found" mentions might be positive or negated mentions that physicians regard as irrelevant information to patient's illness or that are lack of physician's input.

Table 4. Comparison of frequencies of medical problems mentioned in diagnosis and family history.

		Medical Problems Mentioned in Diagnosis		
		Positive	Negated	Not Found
Medical Problems Mentioned in Family History	Positive	58,463	2,887	427,306
	Negated	11,141	1,348	111,649
	Not Found	1,617,480	153,044	-

Table 5. Comparison of the most frequent medical problems in diagnosis and family history.

(A) Positive medical problems in diagnosis while negated in family history		(B) Negated medical problems in diagnosis while positive in family history	
Medical Problem	Frequency	Medical Problem	Frequency
Hypertension	780	Breast Cancer	181
Depression	751	Hypertension	167
Diabetes	539	Depression	119
Cancer	447	Cancer	116
Asthma	391	Asthma	99
Coronary artery disease	378	Stoke	96
Colon cancer	280	Coronary artery disease	96
Hyperlipidemia	178	Diabetes	88
Breast cancer	173	Obstructive sleep apnea	81
Attention deficit disorder	160	Sleep apnea	68
Prostate cancer	160	Attention deficit disorder	58
Anxiety	158	Myocardial infarction	54
Headaches	157	Allergies	53
Seizure	145	Myocardial ischemia	47
Skin cancer	131	Anxiety	34
Osteoporosis	129	Restless legs	34
Migraine headaches	124	Colon cancer	31
Sleep apnea	123	Heart Disease	31
Melanoma	119	Headaches	26
Pain	115	Diarrhea	25

In order to show whether the diagnosed medical problems are mentioned in family history prior to the diagnosis date, we extracted the patients that had the identical medical problems in family history and diagnosis, and calculated the number of years between the diagnosis date of a medical problem and the first date of that medical problem mentioned in patient’s family history. The results for five most common medical problems, hypertension, hyperlipidemia, depression, asthma and cancer, are summarized in Figure 5. We observed that those medical problems were mentioned in the family history up to 15 years prior to the diagnosis date. 36.3%, 33.1%, 39.2%, 25.8% and 52.2% of patients had family history of hypertension, hyperlipidemia, depression, asthma and cancer before they were diagnosed with those medical problems, respectively, and 3.6%, 2.7%, 2.2%, 1.9% and 6.1% of the patients had family history of those problems 10~15 years prior to the diagnosis date.

To show personalized association between family history and diagnosis, we took two patients as examples to illustrate their medical problems in family history and diagnosis. Figure 6 displays the medical problems in their family history and diagnosis associated with the timeline since the first clinical note. For Patient 1, it is clearly shown that asthma and obesity were found in family history about one year before the first diagnosis, and hypertension more than two years before the first diagnosis. Hyperlipidemia was found in family history after the first diagnosis. Pharyngitis, URI, chest pain, abdominal pain and sinusitis were not found in family history because these were specific problems of which the information was not compiled in family history. For Patient 2, hypertension occurred in family history around 6 years before diagnosis and obesity occurred slightly earlier than diagnosis. These results also imply the prediction power of family history for diagnosis.

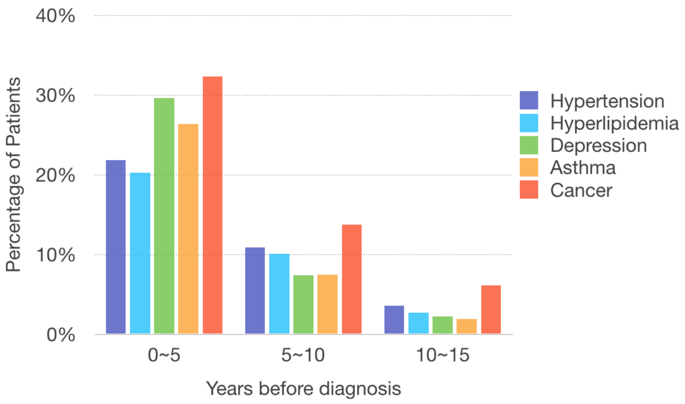
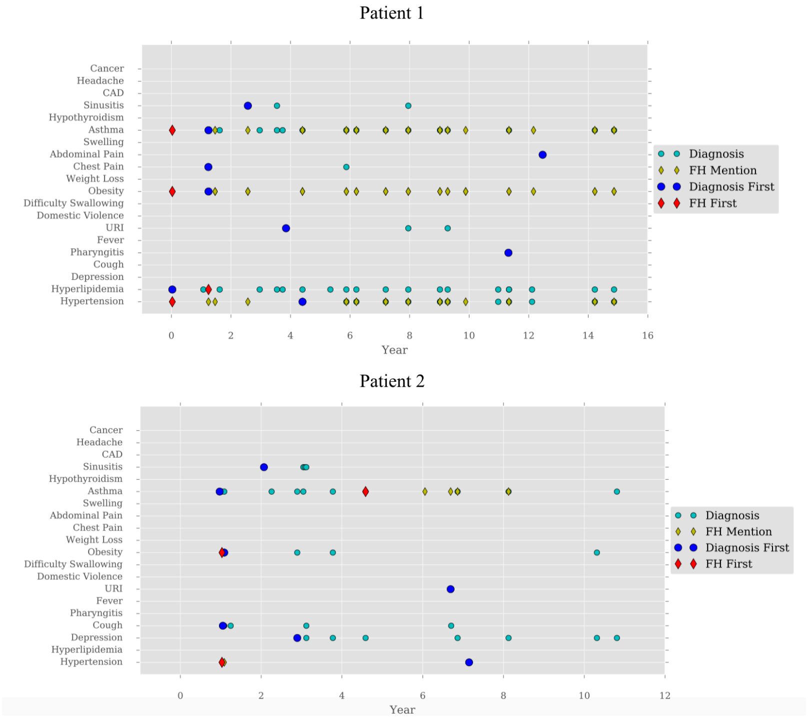


Figure 5. Number of years between the diagnosis date of a medical problem and the first date that medical problem mentioned in patient's family history.


Patient 1

Patient 2

Figure 6. Timeline charts of two patient cases. X-axis represents the timeline of clinical notes while y-axis the medical problems. A turquoise dot represents a medical problem in diagnosis and a blue dot indicates the first date such medical problem is diagnosed. A yellow diamond represents a medical problem in family history and a red diamond indicates the first date such medical problem occurred in family history.

Discussion

We have described a systematic analysis of family history information using a cohort of patients receiving their primary and specialty care at Mayo Clinic. We applied NLP to extract medical problems and family members from the free-text family history. We did not distinguish "maternal relative" and "paternal relative" in the analysis in spite of the importance of specification of side of family for familial disease study. The reason is that extraction of simple family member terms results in a higher accuracy. Future work would consider involving extraction of "maternal" and "paternal" information. In addition, this study focuses on analysis of unstructured free-text family history. A comparison of structured family history and unstructured free-text family history is also interesting and subject to a future study.

Semi-structured family history usually follows certain structures that are frequently used in clinical notes. For narrative family history, physicians spend time gathering relevant family history information, which can be more informative in clinical care. A comparison of semi-structured family history and narrative family history in supporting clinical decision-making would be of interest in a future study.

From our study, we observe that certain diseases are recorded in the family history while others are not. The reason is that recording the family history is highly influenced by the clinical context. For example, the patients with a specific familial condition will be asked by physicians about the relevant family history. Therefore, some rare diseases are included in the family history and some are not. What disease information should be considered and collected in the family history is still a challenge and needs future studies.

The study of agreement between medical problems in family history and those in diagnosis shows some evidence of using family history to predict a patient's future health. Many researchers have found that it is possible to predict medical problems by joint use of family history and other factors [30,31,35–37]. However, few studies utilize the free-text family history from EHR to predict a patient's future health. An automatic system that utilizes NLP tools to extract information from family history and applies probabilistic models to calculate the probability of a patient's future illness is our future study focus. In addition, a timeline visualization tool for showing the information in family history and diagnosis might also facilitate personalized health care, which requires further study. Note that we used exact matches in assessing agreement and did not take into consideration association among the medical problems. One of the future directions would be incorporating the association information leveraging ontologies or empirical data into our analysis.

Conclusion

Free-text family history contains important and valuable information for physicians and clinicians. This is the first systematic analysis of a large free-text family history data set. The aim of this study is to increase the awareness of importance of family history through analyzing the information contained in the free-text family history. We reported the family members and top medical problems mentioned in the corpus as well as their associations. The analysis of patient's diagnosed medical problems and those problems in family history imply the potential use of family history for predicting medical problems. The results also have implications for physicians' training and learning of family history.

References

1. King RA, Rotter JI, Motulsky AG. *The Genetic Basis of Common Diseases*. Oxford university press; 2002.
2. Williams RR, Hunt SC, Heiss G, et al. Usefulness of cardiovascular family history data for population-based preventive medicine and medical research (the Health Family Tree Study and the NHLBI Family Heart Study). *Am J Cardiol*. 2001;87(2):129-135.
3. Kluijtmans LA, Van den Heuvel LP, Boers GH, et al. Molecular genetic analysis in mild hyperhomocysteinemia: a common mutation in the methylenetetrahydrofolate reductase gene is a genetic risk factor for cardiovascular disease. *Am J Hum Genet*. 1996;58(1):35.
4. Loring JF, Wen X, Lee JM, Seilhamer J, Somogyi R. A gene expression profile of Alzheimer's disease. *DNA Cell Biol*. 2001;20(11):683-695.
5. Van't Veer LJ, Dai H, Van De Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(6871):530-536.
6. Zhang L, Zhou W, Velculescu VE, et al. Gene expression profiles in normal and cancer cells. *Science (80-)*. 1997;276(5316):1268-1272.
7. Yoon PW, Scheuner MT, Peterson-Oehlke KL, Gwinn M, Faucett A, Khoury MJ. Can family history be used as a tool for public health and preventive medicine? *Genet Med*. 2002;4(4):304-310. doi:10.1097/00125817-200207000-00009.
8. Endevelt R, Goren I, Sela T, Shalev V. Family history intake: a challenge to personalized approaches in health promotion and disease prevention. *Isr J Health Policy Res*. 2015;4:60.
9. Mehrabi S, Wang Y, Ihrke D, Liu H. Exploring Gaps of Family History Documentation in EHR for Precision Medicine - A Case Study of Familial Hypercholesterolemia Ascertainment. *AMIA Summits Transl Sci Proc*. 2016;2016:160-166. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001769/.
10. Scheuner MT, Wang S, Raffel LJ, Larabell SK, Rotter JI. Family history: a comprehensive genetic risk assessment method for the chronic conditions of adulthood. *Am J Med Genet*. 1997;71(3):315-324.
11. Steinberg GD, Carter BS, Beaty TH, Childs B, Walsh PC. Family history and the risk of prostate cancer. *Prostate*. 1990;17(4):337-347.
12. Hunt SC, Williams RR, Barlow GK. A comparison of positive family history definitions for defining risk of future disease. *J Chronic Dis*. 1986;39(10):809-821.
13. Chen ES, Carter EW, Winden TJ, Sarkar IN, Wang Y, Melton GB. Multi-source development of an integrated model for family health history. *J Am Med Inf Assoc*. 2014:67-80. doi:10.1136/amiajnl-2014-003092.

14. Bill R, Pakhomov S, Chen ES, Winden TJ, Carter EW, Melton GB. Automated extraction of family history information from clinical notes. In: *AMIA Annual Symposium Proceedings*. Vol 2014. American Medical Informatics Association; 2014:1709.
15. Chen ES, Melton GB, Wasserman RC, Rosenau PT, Howard DB, Sarkar IN. Mining and Visualizing Family History Associations in the Electronic Health Record: A Case Study for Pediatric Asthma. In: *AMIA Annual Symposium Proceedings*. Vol 2015. American Medical Informatics Association; 2015:396.
16. 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. doi:86427 [pii].
17. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. *AMIA Annu Symp Proc*. 2008:247-251. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2656021&tool=pmcentrez&rendertype=abstract.
18. Mehrabi S, Krishnan A, Roch AM, et al. Identification of Patients with Family History of Pancreatic Cancer--Investigation of an NLP System Portability. *Stud Health Technol Inform*. 2015;216:604-608. http://europepmc.org/abstract/MED/26262122.
19. Marder K, Levy G, Louis ED, et al. Accuracy of family history data on Parkinson's disease. *Neurology*. 2003;61(1):18-23.
20. Silverman JM, Keefe RSE, Mohs RC, Davis KL. A Study of the Reliability of the Family History Method in Genetic Sudies of Alzheimer Disease. *Alzheimer Dis Assoc Disord*. 1989;3(4):218-223.
21. 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*. 2012;2012:85-92. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3540518&tool=pmcentrez&rendertype=abstract.
22. Liu H, Bielinski SJ, Sohn S, et al. An Information Extraction Framework for Cohort Identification Using Electronic Health Records . *AMIA Summits Transl Sci Proc*. 2013;2013:149-153. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3845757/.
23. Bar-Joseph Z, Gifford DK, Jaakkola TS. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*. 2001;17(suppl 1):S22-S29.
24. Hall JA, Roter DL. Patient gender and communication with physicians: results of a community-based study. *Women's Heal (Hillsdale, NJ)*. 1994;1(1):77-95.
25. Baker DL, Schuette JL, Uhlmann WR. A guide to genetic counseling. 1998.
26. LU IL, TIO A. Family history: the three-generation pedigree. *Am Acad Fam Physicians*. 2005;72:441-448.
27. Yoon PW, Scheuner MT, Peterson-Oehlke KL, Gwinn M, Faucett A, Khoury MJ. Can family history be used as a tool for public health and preventive medicine? *Genet Med*. 2002;4(4):304-310.
28. Pharoah PDP, Day NE, Duffy S, Easton DF, Ponder BAJ. Family history and the risk of breast cancer: a systematic review and meta-analysis. *Int J cancer*. 1997;71(5):800-809.
29. Lothe RA, Peltomäki P, Meling GI, et al. Genomic instability in colorectal cancer: relationship to clinicopathological variables and family history. *Cancer Res*. 1993;53(24):5849-5852.
30. Grant BF, Dawson DA, Stinson FS, Chou PS, Kay W, Pickering R. The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADIS-IV): reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample. *Drug Alcohol Depend*. 2003;71(1):7-16.
31. Milne BJ, Caspi A, Harrington H, Poulton R, Rutter M, Moffitt TE. Predictive Value of Family History on Severity of Illness: The Case for Depression, Anxiety, Alcohol Dependence, and Drug Dependence. *Arch Gen Psychiatry*. 2009;66(7):738-747. doi:10.1001/archgenpsychiatry.2009.55.
32. Oberle I, Camerino G, Heilig R, et al. Genetic screening for hemophilia A (classic hemophilia) with a polymorphic DNA probe. *N Engl J Med*. 1985;312(11):682-686.
33. Castelli WP, Anderson K. A population at risk: prevalence of high cholesterol levels in hypertensive patients in the Framingham Study. *Am J Med*. 1986;80(2):23-32.
34. Gallagher EJ, LeRoith D. Diabetes, cancer, and metformin: connections of metabolism and cell proliferation. *Ann N Y Acad Sci*. 2011;1243(1):54-68.
35. Lalloo F, Varley J, Ellis D, et al. Prediction of pathogenic mutations in patients with early-onset breast cancer by family history. *Lancet*. 2003;361(9363):1101-1102.
36. Kulig M, Bergmann R, Niggemann B, Burow G, Wahn U, Group MASS. Prediction of sensitization to inhalant allergens in childhood: evaluating family history, atopic dermatitis and sensitization to food allergens. *Clin Exp Allergy*. 1998;28:1397-1403.
37. Hansen LG, Halken S, Høst A, Møller K, Østerballe O. Prediction of allergy from family history and cord blood IgE levels. *Pediatr Allergy Immunol*. 1993;4(1):34-40.