Exploring Gaps of Family History Documentation in EHR for Precision Medicine - A Case Study of Familial Hypercholesterolemia Ascertainment

Saeed Mehrabi, PhD, Yanshan Wang, PhD, Donna Ihrke, Hongfang Liu, PhD
Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA; USA

Abstract

In the era of precision medicine, accurately identifying familial conditions is crucial for providing target treatment. However, it is challenging to identify familial conditions without detailed family history information. In this work, we studied the documentation of family history of premature cardiovascular disease and hypercholesterolemia. The information on patients’ family history of stroke within the Patient-provided information (PPI) forms was compared with the information gathered by clinicians in clinical notes. The agreement between PPI and clinical notes on absence of family history information in PPI was substantially higher compared to presence of family history.

Introduction

It has been shown that a wide range of adult conditions such as diabetes, cardiovascular diseases, Alzheimer’s and cancers have hereditary roots. Accurate family history information can be very helpful in precision medicine that tailors the treatment to the individual characteristics of patients. For instance, the risk of having colon cancer for individuals with family history of colon cancer is two fold, which makes individuals with positive family history of colon cancer the best candidates for genetic testing and preventive screening.

The family history information can be available in clinical notes by “documenting parents’ and siblings’ age and health (or age and cause of death), as well as a checklist of conditions with environmental and hereditary etiologies.” Although most family practice training programs emphasize and educate practitioners on documenting family history information, direct observation of family history is normally completed on patients’ first visit and once it is recorded in patients’ medical record, its discussion would be less likely in the subsequent visits. However family history is not static and a more accurate and comprehensive information should be collected periodically. The family history information can also be available through self-administered medical/family history questionnaires, which have long been encouraged in addition to conventional patient interviews in ambulatory care settings. Benefits such as time saving for healthcare providers, follow-up and identification of additional information on problems documented in the questionnaires, and more complete patient record have been associated with addition of questionnaires to patient examination.

At Mayo Clinic, family history information can be obtained through either clinical notes or patient provided information (PPI) questionnaires. PPI collects information on patients’ prior family history, current visit information, allergies, medications, substance use history, etc. in an electronic format. The system uses a set of decision rules to determine if any additional forms such as asthma management questions, depression screening and treatment monitoring are required. These forms can be completed electronically through Mayo Clinic’s patient portal or mailed to patient in paper format. The completed paper forms are scanned using a large-scale distributed network of smart scanners. The PPI system is electronic health records (EHR) agnostic and can be seamlessly integrated into clinical notes.

In our effort of screening patients with familial conditions, multiple challenges to recognize familial conditions based on EHRs were identified and significant gaps in between clinical notes and PPI were discovered. First, family history information may only be available in free text format. In order to obtain such information, natural language processing (NLP) is needed. In Mayo EHR, we observed that family history information in clinical notes could be in various formats: short sentences, narratives, tables or mixture of above. A complex NLP system is needed in order to accurately extract family history information. Secondly, family history information may be available in structured format but not in the granularity needed (parental or maternal side) or lack the required information such as age of onset of the conditions.
In this work, we study the documentation of family history for the feasibility of automated Familial Hypercholesterolemia (FH) ascertainment from EHRs and illustrate those challenges.

Related Works

Review of various tools such as self-completed family history questionnaire, physician administered electronic questionnaire, automated family history telephone interview, etc. used in primary care setting for collection of familial cancer history showed 46-78% improvement in data recording compared to family history recording in patients’ charts and 75 to 100% agreement with genetic interviews.

St. Sauver et al studied the agreement between information from patient/family history questionnaires with Mayo Clinic Medical diagnostic index on cardiovascular disease (CVD) and CVD risk factors (i.e. blood pressure, cholesterol, triglycerides, etc.) for 161. They found substantial agreement between patients report and medical index on the absence of CVD conditions ranging from 90% for high cholesterol to 98% for medical problem or surgery. The positive agreement values, however, ranged from 31% for medical problem or surgery to 78% for high blood pressure. Dhimen et al also studied the availability and quality of family history of Congestive Heart Failure (CHD) using the health improvement network database containing data from 537 primary care practices in the UK. They used multilevel logistic regression analysis to assess the availability and quality of documented family history information. Presence of family history of CHD was recorded in 9.3% of their total cohort of 1,504,535 patients. Patients aged 50-59 had higher odds of having their family history recorded compared to patients aged 20-29, however the level of recording fell in patients aged above 70. It has been shown that the quality of patient reported information and consequently the level of agreement between questionnaires and medical notes largely depends on the structure of the questionnaire, nature of the disease, questions asked, and characteristics of the individual who is completing the questionnaire.

The previous mentioned studies used coded data to compare patient reported information with clinical notes ignoring the information documented in patients’ reports in free text format. Pakhomov et al used NLP to determine the agreement between patient reported symptoms of chest pain, dyspnea and cough and their documentation in EHR. They used both EHR and PPI data of Mayo Clinic patients during January 1st to June 30th of 2006 consisting of 121,891 patients and randomly selected 1,119 patients with positive mentions of the three symptoms. The positive/negative agreement for chest pain, dyspnea and cough was 74/78 %, 70/76 %, and 63/75 % respectively. Positive agreement was lower for women and younger patients while negative agreement was lower for men and older patients across the three symptoms.

Background of Familial Hypercholesterolemia (FH)

FH is a genetic disorder where the body is unable to remove low-density lipoprotein (LDL) from blood, which eventually results in having high levels of LDL. FH increases the risk of premature ASCVD that can be easily treated if detected early. FH is a relatively common genetic disorder with reported prevalence of as high as 1 in 137. It is reported that one-fifth of patients with CHD younger than 45 years have heterozygous form of low-density lipoprotein receptor (LDLR) mutation causing FH. In individuals with heterozygous FH (heFH), CHD occurs in early middle age at about 35 years if untreated, if treated since age 18 this threshold shifts towards age 48, and treatment from age 10 postpones the onset of CHD to age 53. Young adults aged 20-39 years have a 100-fold FH-associated increase in the risk of death from ASCVD and approximately a 10-fold increase in overall mortality rates compared with the age-and sex-matched general population. In heFH patients life expectancy is shortened by 20–30 years with sudden death and myocardial infarction as the principal causes of mortality. Nearly 50% of men with heFH experience ASCVD by age 50, and 30% of women by age 60, thus increasing the healthcare and economic burden of disability and dependency. Over 80% of patients with FH in western countries are undetected, and the proportion of undetected cases in the US may be greater than 95%. Recently, a national registry-based initiative for improving FH detection and treatment has been proposed. However, there are no systematic approaches for identifying FH patients or screening their relatives in the US. In order to identify FH patients based on EHR according to the diagnosis criteria defined by Dutch Lipid Clinic Network (DLCN), there is a need for accurate identification of family history of premature ASCVD and hypercholesterolemia.
Materials and Methods

Data set

Employee and Community Health (ECH) primary care population is a cohort of 138k patients at Mayo Clinic. We selected any patient who had hypercholesterolemia (n=10,785) or mentions of LDL measurement in their unstructured clinical notes (n=26,767) that created a cohort of 27,998 unique patients. Patients with LDL measurements of above 190 in their structured lab reports added an extra of 3,216 patients (n=31,214). We excluded 1,938 patients for which we did not have research permission reducing the number of total patients to 29,276. Mayo Clinic Enterprise Data Trust (EDT) was used to retrieve PPI for the 29,276 selected patients. EDT is a clinical repository that collects its data from multiple sources within Mayo Clinic and can be utilized to efficiently search millions of patient records including diagnosis, laboratory, clinical notes, PPI, etc. Table 1 shows examples of questions related to family history of stroke or transient ischemic attack (TIA), high cholesterol, and heart attack retrieved from various PPI forms.

| Disease          | Questions                                   |
|------------------|---------------------------------------------|
| Stroke/TIA       | Relatives with - Strokes                    |
|                  | Which blood relatives have had Stroke/TIA?  |
|                  | Have you or a family member had a stroke?   |
| High Cholesterol | Relatives with High Cholesterol             |
|                  | Which blood relatives have had high cholesterol? |
|                  | Have you or a family member had high cholesterol? |
| Heart Attack     | Have your relatives had a heart attack?     |

Methods

Clinical terms pertaining to family history of heart disease, stroke/TIA, and high cholesterol were assembled by our medical expert team, which were expanded by manual review of clinical notes and UMLS Metathesaurus. Regular expression patterns were developed to identify the assembled medical terms within the clinical text. We used our previously developed family history identification method incorporated with MedTagger for section detection, sentence tokenization, and negation detection. MedTagger is a suite of tools including sectionizer adapted from SecTag to detect sections, a rule-based context annotator adapted from ConText assigning each concept a status modifier (i.e., positive, negative, and probable), and an experiencer modifier to specify if the patient or other family members (i.e., father, mother, etc.) have the findings. Section segmentation in clinical notes is an important first step in family history that can be challenging, as no fixed terminology is available for section headers. However Mayo Clinic records are CDA 1.0 compliant with codified sections. In order to associate multiple family members and diagnoses in a sentence, our previously developed rule based algorithm was utilized.

Evaluation

Equations (1) and (2) were used to assess the agreement between PPI and NLP results.

Positive Agreement= \frac{2a}{N+(a-d)} \tag{1}

Negative Agreement= \frac{2d}{N-(a-d)} \tag{2}

Where \(a\) = presence of family history in both PPI and medical records, \(b\) = presence of family history in medical records and its absence in PPI forms, \(c\) = absence of family history in medical records and its presence in PPI forms, \(d\) = absence of family history in both medical records and PPI forms, and \(N\) is the total number of observations \((N=a+b+c+d)\). If a patient had reported two family members (mother and grandparent) with stroke in PPI form while only one family member (mother) with stroke were documented in clinical notes, “a” would be equal to one due to presence of mother in both PPI and medical notes and “c” would be one because of the presence of grandparent in PPI and its absence in clinical notes.
These measurements have been proposed to evaluate the agreement between patient reported information and the medical records instead of traditional kappa values because statistics based on kappa are dependent of imbalances in marginal totals and may not provide an accurate estimation of agreement between patient reported information and information provided in the medical records.\textsuperscript{10,14}

Results

Out of the total number of 29,276 patients present in our study, NLP identified 26,441 (90.3\%) patients with family history information in clinical notes, while 29,220 (99.8\%) of them had family history information in their PPI form. 23,942 of the patients had family history information in both PPI and clinical notes, 2,826 had only information in PPI and 47 had only information in medical records. Of the 26,441 patients with family history information in their medical reports, 22,487 (85\%) had information about the affected family member and 11854 (44.8\%) had both age and family member information.

Family history of stroke/TIA was used to evaluate the agreement between PPI and medical records. We randomly selected 400 patients with family history information in their PPI and manually checked the information regarding family history of stroke in their PPI forms and family history of stroke extracted from their clinical notes based on NLP. Table 2 shows the confusion table regarding the agreement between PPI forms and medical records.

Table 2. Positive and negative agreement measurements of stroke mentions in PPI forms and medical records

| Patient Provided Information | Presence | Absence |
|-----------------------------|----------|---------|
| Medical Records             | Presence | a=14    | b=26    |
|                             | Absence  | c=17    | d=345   |

The positive and negative agreements are 0.3994 and 0.9413 respectively.

Out of 26 family member instances of stroke/TIA present in clinical notes and absent in PPI forms, 8 of the instances were due to differences in the family members listed in PPI forms and 18 were patients without any information in PPI forms. From the total of 17 family member instances present in PPI and absent in clinical notes only one was due to differences in the relative listed (i.e. Brother with stroke in PPI and grandparent in clinical notes) and there was no information on family history of stroke for the other 16 patients. The agreement between PPI and clinical notes on absence of family history information is much higher compared to presence of information in PPI forms in regard to family history of stroke/TIA (94\% compared to 40\%).

It should be noted that the NLP results are hundred percent correct. Evaluation of NLP performance on a random set of 100 patients resulted in accuracy of 81\% (19 inaccurate family member/negation association). Majority of these errors (15 out of 19) were due to incorrect family member associations while only 3 accounted for negation detection errors and in one case both family association and negation were incorrect.

Discussion

Similar to a previous study that investigated the agreement between PPI and medical reports on CAD and its risk factor,\textsuperscript{10} we found a very strong agreement on negative responses. Positive agreement on the other hand was much lower with 40\% of patients with a positive response to the question of relatives with stroke on their PPI forms had no family history information in their clinical notes. However, the two sources complement rather than contradict each other. Although union of the information represented in PPI and clinical records can create a comprehensive family history records, still important details such as age of disease onset and affected relatives are largely missing from both sources. Such missing information limits the application of risk assessment tools, for instance if a patient had a mother with colon cancer at age 50 or above he has a two-fold increase in risk of having colon cancer, however if he had a mother who had colon cancer at age 40 there is much higher risk (50-100\%) of getting cancer due to possible inheritance of her genes. Also environmental factors and life style of family members are very
important information that are missing. Knowledge regarding family members’ lifestyle habits such as smoking and alcohol consumption is crucial to rule out environmental versus genetic factors causing heart disease or cancer at younger age than 50. The size of the family also matters in determining familial conditions. If there are more family members, there is a higher chance to observe familial conditions.

We faced some challenges in extraction of family history information from clinical notes using our NLP system such as differentiating between current age versus age of disease onset (i.e. Mother is 83 years old and had a CABG at age 77), associating family members with their age when multiple family members with age information were in one sentence (i.e. Remarkable for a father who died in his 70's of coronary artery disease and a mother who passed away in her 80's with a similar disorder.), implicit representation of age (no family history of early stroke/tia), co-referencing when disease (high cholesterol) and family member (mother) were in two separate sentences (Mother was adopted. No known early heart attacks, kidney problems, or high cholesterol requiring medication.), etc.

PPI forms have also certain restriction compared to clinical notes such as lack of a dedicated section for patients to document information regarding the age of disease onset or paternal or maternal side of their relatives.

Conclusion and Future Works

Illustrated through family history of premature CHD that is one of the criteria for diagnosis of heFH in adults based on Dutch Lipid Clinic Network, we have identified that there are significant gaps in the family history documentation. Positive family history is the basis for the diagnosis of many familial conditions and with the increase of precision medicine practice, it can not only be a great source of information for diagnosis and screening but also can be utilized for targeted treatment. And therefore adequate family history information including detailed information such as affected family members and their social history, age of disease onset, and specific information regarding the disease in question are crucial.

In genetic testing, up to three generations of family history is required to create a pedigree. However patients’ recall of family history is often inaccurate and inadequate. Family events and gatherings such as Christmas, Thanksgiving, etc. are best opportunity to inquire about the family history from the older generation in the family. The Surgeon General declared the Thanksgiving to be the national family history day since 2004 to encourage people to talk about and write down their family health history.

The case study shows that majority of the patients do have family history information available in EHR but not detailed enough for automated familial condition ascertainment. The study also illustrates the diversity in family history documentation in clinical notes requiring more complex NLP algorithms. Future study will include the survey of common familial conditions and identification of associated family history information. Additional future work will focus on the development of automated familial condition ascertainment approaches that can tolerate missing and incomplete family history information.

References

1. Wilson BJ, Qureshi N, Santaguida P, Little J, Carroll JC, Allanson J, et. al. Systematic Review: Family History in Risk Assessment for Common Diseases. Ann Intern Med 2009, 151 (12), 878–885.
2. 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, 304 –3310.
3. Degowin EL, Degowin RL. Bedside Diagnostic Examination, 2nd ed.; Macmillan Co: New York, NY, 1969.
4. Medalie JH, Zyzanski SJ, Langa D, Stange KC. The family in family practice: is it a reality? J Fam Pract 1998, 46 (5), 390-396.
5. Acheson LS, Wiesner GL, Zyzanski SJ. Family history-taking in community family practice: implications for genetic screening. Genet Med 2000, 2, 180 –1185.
6. Gumpel JM, Mason AM. Self-administered clinical questionnaire for outpatients. Br Med J 1974, 2, 209-212.
7. Inui TS, Jared RA, Carter WB, Plorde DS, Pecoraro RE, Chen MS, et. al. Effects of a self-administered health history on new-patient visits in a general medical clinic. Med Care 1979, 17, 1221–1228.

8. Eton DT, Beebe TJ, Hagen PT, Halyard MY, Montori VM, Naessens J, et. al. Harmonizing and consolidating the measurement of patient-reported information at health care institutions: a position statement of the Mayo Clinic. Patient Related Outcome Measures 2014, 5, 7–15.

9. Qureshi N, Carroll JC, Wilson B, Santaguida P, Allanson J, Brouwers M, et. al. The current state of cancer family history collection tools in primary care: a systematic review. Genet Med 2009, 11(7), 495-506.

10. St Sauver JL, Hagen PT, Cha SS, Bagniewski SM. Agreement between patient reports of cardiovascular disease and patient medical records. Mayo Clin Proc 2005, 80 (2), 203–210.

11. Dhiman, P, Kai J, Horsfall L, Walters K, Qureshi N. Availability and quality of coronary heart disease family history in primary care medical records: implications for cardiovascular risk assessment. PLoS One 2014, 9 (1), e81998.

12. Haapanen N, Miilunpalo S, Pasanen M, Oja P, Vuori I. Agreement between questionnaire data and medical records of chronic diseases in middle-aged and elderly Finnish men and women. Am J Epidemiol 1997, 145 (8), 762-769.

13. Colditz GA, Martin P, Stampfer MJ, Willett WC, Sampson L, Rosner B, et. al. Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. Am J Epidemiol 1986, 123 (5), 894-900.

14. Pakhomov SV, Jacobsen SJ, Chute CG, Roger VL. Agreement between patient-reported symptoms and their documentation in the medical record. Am J Manag Care 2008, 14 (8), 530-539.

15. Horton JD, Cohen JC, Hobbs HH. PCSK9: A convertase that coordinates LDL catabolism. J Lipid Res 2009, 50 Suppl, 172-177.

16. Huigen R, Hutten BA, Kindt II, Vissers MN, Kastelein JJ. Discriminative ability of LDL-cholesterol to identify patients with familial hypercholesterolemia: A cross-sectional study in 26,406 individuals tested for genetic FH. Circ Cardiovasc Genet 2012, 5, 354-359.

17. Starr BB, Hadfield SG, Hutten BA, Lansberg PJ, Leren TP, Damgaard D, et. al. Development of sensitive and specific age- and gender-specific low-density lipoprotein cholesterol cutoffs for diagnosis of first-degree relatives with familial hypercholesterolaemia in cascade testing. Clin Chem Lab Med. 2008, 46, 791-803.

18. Scientific Steering Committee on behalf of the Simon Broome Register Group. Mortality in treated heterozygous familial hypercholesterolaemia: Implications for clinical management. Atherosclerosis. 1999, 142, 105-112.

19. Alonso R, Mata P, Zambon D, Mata N, Fuentes-Jimenez F. Early diagnosis and treatment of familial hypercholesterolemia: Improving patient outcomes. Expert Rev Cardiovasc Ther 2013, 11, 327-342.

20. Austin MA, Hutter CM, Zimmern RL, Humphries SE. Familial hypercholesterolemia and coronary heart disease: A huge association review. Am J Epidemiol 2004, 160, 421-429.

21. NICE. Identification and management of familial hypercholesterolaemia. ; National Institue for Health and Care Excellence, 2008.

22. Civeira F. Guidelines for the diagnosis and management of heterozygous familial hypercholesterolemia. Atherosclerosis 2004, 173, 55-68.

23. Nordestgaard BG, Chapman MJ, Humphries SE, Ginsberg HN, Masana L, Descamps OS, et. al. Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: Guidance for clinicians to prevent coronary heart disease: Consensus statement of the european atherosclerosis society. Eur Heart J 2013, 34, 3478-3490.

24. O'Brien EC, Roe MT, Fraulo ES, Peterson ED, Ballantyne CM, Genest J, et. al. Rationale and design of the familial hypercholesterolemia foundation cascade screening for awareness and detection of familial hypercholesterolemia registry. Am Heart J 2014, 167 (3), 342-349.
25. World Health Organization. Familial hypercholesterolemia—report of a second WHO Consultation; World Health Organization: Geneva, Switzerland, 1999.

26. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc 2010, 17 (2), 131–135.

27. Mehrabi S, Krishnan A, Roch AM, Schmidt H, Li D, Kesterson J, et. al. Identification of Patients with Family History of Pancreatic Cancer - Investigation of an NLP system Portability. Stud Health Technol Inform. 2015;216:604-8.

28. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et. al. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc, 2013; pp 149–153.

29. Versmissen J, Oosterveer DM, Yazdanpanah M, Defesche JC, Basart DC, Liem AH, et. al. Efficacy of statins in familial hypercholesterolaemia: A long term cohort study. BMJ 2008, 337, a2423.

30. Norata GD, Ballantyne CM, Catapano AL. New therapeutic principles in dyslipidaemia: Focus on LDL and Lp(a) lowering drugs.. Eur Heart J 2013, 34, 1783 -1789.

31. Hayflick SJ, Eiff MP, Carpenter L, Steinberger J. Primary care physicians’ utilization and perceptions of genetics services. Genet Med 1998, 1 (1), 13-21.

32. Medalie JH, Zyzanski SJ, Langa D, Stange KC. The family in family practice: is it a reality? J Fam Pract 1998, 46 (5), 390-396.