Enhancing FHx collection and documentation with a chatbot and NLP pipeline

Michelle Hoang Nguyen[1][0000-0002-7689-646X]
[1]Johns Hopkins School of Medicine, 21205, USA
mnguye79@jhmi.edu

Abstract. Although family history (FHx) contains vital information for genetic risk assessment, few patients have complete FHx in their electronic health record (EHR), often due to provider barriers to collect this information in clinic [1-3]. Asynchronous digital health strategies that rely on patient-entered family history may be able to reduce clinician burden and improve FHx completion rate by empowering patients to electronically complete their FHx at a time and place that is convenient for them. However, electronic FHx forms do not usually provide guidance for the patient to collect their FHx, and therefore, may result in lower quality FHx data limiting its ultimate clinical utility. Through our preliminary work developing a patient-facing chatbot-based FHx collection tool, our hypothesis is that providing patients with their previous FHx data and educational resources in a chatbot-guided FHx collection format will improve FHx data quality and data utility for clinicians. To address the challenge of incomplete FHx present in the electronic health record (EHR), we propose developing a FHx collection pipeline to supplement FHx in the EHR with chatbot-collected FHx by completing two aims. In Aim 1, we will define the data quality of structured EHR-derived FHx data and assess natural language processing (NLP) strategies to unlock potential from unstructured FHx data. In Aim 2, we will define FHx report and data collection improvements for patient-centered FHx data collection with a large language model (LLM)-augmented chatbot.

Keywords: family health history, information extraction, electronic health record, natural language processing, chatbot, digital health, large language models

1 Problem space

1.1 FHx is an important genetic risk predictor that is not often found in complete and computable formats in the EHR, limiting clinical risk assessment.

Although FHx is widely recognized as a crucial step of genetic risk assessment, less than 50% of U.S. adults report actively collecting their FHx [4]. Most patients do not have complete FHx in their electronic health record (EHR) [5]. FHx is an important genetic and environmental indicator of risk and is included in risk algorithms for common chronic conditions such as diabetes type 2 and cardiovascular disease[6]. Therefore, if key FHx fields are missing, clinicians may have an inaccurate picture of a patient's health and underestimate a patient's risk for disease [5,7,8]. These problems may be greater for patients who are members of demographic minorities that may be even less likely to have complete FHx, leading to further treatment and outcome disparities [9].

1.2 Patient and provider barriers have limited the collection of high quality FHx data.

Barriers to collecting FHx have persisted for both patients and providers. Providers face challenges to collect FHx due to a lack of time during patient visits and limitations of current tools to document FHx comprehensively [1–3]. Patient challenges have included a lack of awareness of FHx's value to assess disease risk, great effort required to confirm conditions present among family members, the need to regularly update FHx to maintain accuracy, and insufficient clinician encouragement to complete FHx documentation [1,4,10,11].

2 Goal and research questions

The proposed pipeline will use an innovative approach to integrate EHR-derived data with high-quality patient-entered FHx data, through use of an LLM-enabled interactive chatbot. The informatics benefits of this work are to generate insights on how to develop LLM-enabled chatbots to improve FHx data collection and ultimate data utility for clinicians, which has not been thoroughly explored in the health informatics community. The clinical benefits of this work lie in its ability to address incomplete and/or missing FHx data in the EHR. This missingness may be more prevalent within patient groups belonging to demographic minorities, which leads to inequitable genetic risk assessment and care delivery for many common chronic conditions. By improving the process of FHx data collection with a guided, informative chatbot, we believe that this tool may lower the barriers for patients and their families to collect complete FHx data of high clinical utility, which would enable better care delivery downstream.
We will accomplish this goal and answer the following questions by completing two specific aims:

2.1 Aim 1. What is the current quality and accessibility of FHx data in the EHR? How can we use NLP strategies to make FHx data accessible and structured in a clinically useful format?

Low quality FHx in the EHR limits patients' access to preventative care and genetic screening. We believe that there is untapped potential to utilize FHx captured in the unstructured clinical notes. By using novel natural language processing (NLP) strategies to extract FHx from free-text clinical notes, we assert that we can improve upon structured FHx data completeness, especially for individuals who do not have structured data in the EHR at all. In Aim 1, we will define the data quality of structured EHR-derived FHx data and assess natural language processing (NLP) strategies to unlock potential from unstructured FHx data and create a patient pedigree visualization. The goal for Aim 1 is to assess improvements to FHx data quality by aggregating FHx data from unstructured and structured data sources by benchmarking against a gold standard dataset of patient pedigrees collected by genetic counselors. Achieving Aim 1 of our proposal will satisfy the EHR-derived FHx requirement of our pipeline.

2.2 Aim 2. How can we improve patient facing FHx report format and FHx data collection with LLMs?

Aim 2 will build upon our previous efforts to develop a chatbot for FHx data collection and adopt the pedigree visualization accomplished in Aim 1. In Aim 2, we will define FHx report and data collection improvements for patient-centered FHx data collection with a large language model (LLM)-augmented chatbot. We hypothesize that by leveraging LLM capabilities, our proposed chatbot framework will have improved conversational flow and an improved patient-facing report design that will enable a more guided FHx collection experience. By completing Aim 2, we strive to address patient-facing barriers to collect FHx that are remain unsupported by current FHx e-forms.

3 Proposed methods

3.1 Aim 1

Aim 1A. Structured family history analysis – determine baseline characteristics of structured FHx data

The goals of Aim 1A are to a) determine the availability of any structured FHx data for a cohort of patients who have pedigree data recorded by genetic counselors and any demographic characteristics that are associated with data availability, b) examine data completeness of FHx by measuring completeness of important features, such as family member, side of the family, condition, age of onset, c) define data completeness and data concordance by data source (patient or provider). We will perform comparative analyses to identify differences between patients who have structured FHx and those who do not, along demographic features. For patients who do have structured FHx, we will perform additional analyses on FHx data quality as modulated by FHx source (patient vs. provider). Examples of FHx data quality measures will include the following: unique family relation- conditions, unique instances of age of onset, unique FHx comments, and generational level of FHx captured, percent agreement between patient and provider-based FHx (for patients who have both). Aim 1A will lay the foundation on which to assess potential FHx data quality improvements with unstructured data.

Aim 1B. Unstructured family history analysis – determine the performance of NLP system to extract unstructured FHx data

Unstructured clinical notes may hold important FHx information not present in structured fields but is difficult to access by clinicians. To utilize this data, we aim to assess the ability of natural language processing strategies to perform document-level relation-extraction of the family relation-conditions in the unstructured clinical notes. A set of three annotators will annotate a corpus consisting of 5% of the available clinical notes (from the cohort selected in Aim 1A) with overlap on 50% of this corpus. Using a semi-supervised self-training approach, we will develop a graph-based deep learning system with different pre-trained language models to represent text (i.e. BERT [12], RoBERTa [13], Clinical BERT [14]). We will assess the overall performance of this system using precision, recall, F1 score and performance along specified relation-

entities of interest. Aim 1B will transform the latent FHx information from unstructured clinical notes into a usable format to supplement structured information gathered in Aim 1A.

Aim 1C. Harmonization of structured and unstructured FHx - Examine data quality from combining sources and creating patient pedigrees

Our overarching goal in Sub-Aim 1C is to determine our capability to create effective patient pedigrees by combining quality from structured and unstructured FHx data. Therefore, we will first assess data concordance between structured and unstructured FHx sources and then assess resulting data completeness from combining unique entity-relation pairs from both data sources. We will explore the intersections and differences in data across unstructured fields, patient-entered structured FHx data, and provider-entered structured FHx data to identify whether patient-entered FHx data has unique family-member condition relations as compared to both provider-entered structured fields and unstructured data. Then we will assess ultimate data quality improvements in FHx and patient demographic representation by aggregating structured and unstructured data using similar data quality measures as Aim 1A. In Aim 1C, we will use the unique FHx data from both sources to build an electronic pedigree chart to visualize the combined data for each patient.

3.2 Aim 2

Aim 2A. Define and study three design options for a LLM-augmented FHx patient data report

In our previous work assessing the usability of a flow-based FHx chatbot, we collected approximately 200 FHx transcripts for two crowd-sourced cohorts on Amazon Mechanical Turk and Qualtrics Panel. From these de-identified transcripts, we will perform a design study using a GPT-based LLM model to improve the design and utility of the FHx reports.

From the previous transcript auto-generated by the chatbot, we will transform the basic transcript text format into a more readable layout. Additionally, we propose 3 areas of report improvements informed by feedback we received in preliminary work. The modulations will be 1) the formality of tone, 2) structure of tailored response, and 3) graphical representation of FHx summary. Next, we will design prompt engineering strategies for each of these three report improvement areas. We will log our prompt engineering strategies to finetune characteristics of the design including different levels of the characteristics: report tone (formal, informal), structure of tailored response (no tailored responses, open-ended guidance based on data entry, specific talking points for discussion with clinicians and family members), and graphical FHx summary style (table, graph, pedigree). We will design survey questions to solicit opinions and preferences based on modulations of the aforementioned design characteristics for five randomly selected transcripts.

Aim 2B. Assess the agreement between chatbot-collected data and a gold standard pedigree and characterize genetics professionals' opinions on the clinical usefulness of the chatbot-generated summary of an LLM-enabled chatbot to deliver guided patient FHx data collection

Our previous chatbot design will be extended in Aim 2B to include samples of EHR-extracted FHx data (from Aim 1B) to personalize conversations and use large language model capabilities to further support chatbot conversational flow as well as construct a patient-facing report and clinician-facing pedigree (from Aim 1C).

To evaluate the agreement between chatbot-collected data and a gold standard pedigree and characterize genetics professionals' opinions on the clinical usefulness of the chatbot-generated summary, we will conduct a study with a set of ten clinical geneticists and genetic counselors who will each complete chatbot data collection for ten patient case studies to yield an ultimate patient report and clinician-facing pedigree. Along with survey questions developed in Aim 2A regarding report design features, we will design and administer key informant surveys to assess clinical validity (agreement between report pedigree with gold standard pedigree), clinical utility (usefulness of the format of the report pedigree), and patient appropriateness and utility (based on clinician opinions of the chatbot generated report and user experience with the data collection mechanism).

4 Prior work

To address challenges to collect FHx and better understand the benefits and drawbacks between different web-based modalities, we have studied the potential for a flow-based chatbot approach to collect FHx. The primary objective of our preliminary work was to understand the usability of a chatbot as compared to a form-based method and observe any chatbot-specific user interface characteristics that may explain this usability. To do this we created an interactive web-embedded chatbot called KIT to administer a three-generational family history survey. We studied its overall usability, perceived usefulness to collect FHx, and how engaging it was to use, when compared to a form-based method[15].

4.1 Findings from preliminary work.

From our preliminary comparison of a flow-based FHx history chatbot and a standard web-based form, we observed that chatbot users reported higher usability on the System Usability Scale (SUS) [16] than form-based users [15]. We also discovered that the few chatbot-specific features we introduced, such as its personality and conversational nature, may have contributed to a more usable experience. We hypothesize that the higher usability reported by KIT users may in part be due to those features.

Although at baseline chatbot users reported higher usability, based on user comments, there is still room for performance improvements for the chatbot's personality, onboarding process, and its management of errors. We proposed three areas to enhance chatbot use and found that all three (gamification, media elements, and personalization) were highly endorsed with over 50% of the respondents indicating a

high or moderate ranking. The priority ranking of features according to our findings were: #1-personalization, #2-media elements, and #3- gamification.

We compared the perceived usefulness of the form-based FHx data collection and the KIT final summary report for a primary care visit. The summary report from the form was a print view of all FHx values that were entered, and for KIT it was a user transcript. We were motivated to explore FHx report usefulness based on evidence of information gaps for both patients and providers to understand the value of FHx for actionable use in their care and condition management[1.] Although our quantitative findings were positive, qualitative user comments indicated that the reports could be improved in several ways: clearer formatting, purpose, and more value beyond the display of data collected (e.g., summary table, personalized response). In future FHx tool design, it is important to introduce what the use of the final report might be at the start of data collection, to intentionally customize it for consumer use, and to understand patient and provider preferences and attitudes towards features such as personalized FHx risk assessment, which informed the design of our specific aims.

5 Expected results

5.1 Aim 1 – Expected results.

Aim 1A - While preliminary data suggest that most of the patients who have seen specialty genetics providers have at least one unique family relation-condition pair in EHR structured FHx, we expect to discover demographic differences between patients who have and do not structured FHx. Additionally, we hypothesize that there are FHx data quality differences between provider and patient-entered FHx data.

Aim 1B - From this experiment, we expect that a graph-based strategy will encode family-member-condition relations well and have higher performance for relation extraction in comparison to other methods. However, the semi-supervised approach may yield lower training data quality due to potentially inaccurate pseudo-labels and may limit performance. If this is true, we may take a different approach by first training on open-source datasets for FHx relation-extraction such as the 2019 n2c2/OHNLP dataset [17] and then assess performance on our corpus through a transfer learning approach.

Aim 1C - We expect that by harmonizing FHx data from structured and unstructured sources, we will yield a more complete picture of FHx for each patient. Additionally, we hypothesize that unstructured FHx will fill the gaps for patients without structured FHx, and this may mitigate systematic FHx missingness for patients of different demographic groups.

5.2 Aim 2 – Expected results.

Aim 2A – We expect that the original transcript format will be transformed into a more usable report through an LLM-augmented re-design process. Additionally, we hypothesize that some design characteristics may be less costly to redesign with an LLM, with minimal harm. However, tailored response recommendations may be more challenging to implement, as LLMs are prone to hallucinations, which may result in

harmful responses. Because of these considerations, we may build additional guardrails against hallucinations by pre-defining appropriate talking points in our prompt design.

Aim 2B – We expect to reveal generated pedigrees via FHx chatbot data collection that are consistent with gold standard pedigrees. Additionally, we hypothesize that clinicians will have largely positive opinions about the LLM-enabled chatbot to collect FHx data and to generate appropriate post-data-collection outcomes for both patients and clinicians.

6 Areas for feedback

For the outlined proposal, we are seeking feedback related to –

- Assessing appropriateness of proposed statistical analyses/evaluation related to Aims 1 and 2
- Assessing clinician survey study design for Aim 2B

Acknowledgments. Thank you to my advisor, Dr. Casey Overby Taylor, for her guidance in developing this research proposal, as well as to my committee members, Drs. Kadija Ferryman, Ada Hamosh, João Sedoc, and Ayah Zirikly for their input and feedback.

Disclosure of Interests. The author has no competing interests to declare.

References

1. Wildin RS, Messersmith DJ, Houwink EJF. Modernizing family health history: achievable strategies to reduce implementation gaps. J Community Genet. 2021 Jul;12(3):493–496. PMCID: PMC8241955
2. Rich EC, Burke W, Heaton CJ, Haga S, Pinsky L, Short MP, Acheson L. Reconsidering the Family History in Primary Care. J Gen Intern Med. 2004 Mar;19(3):273–280. PMCID: PMC1492151
3. Taber P, Ghani P, Schiffman JD, Kohlmann W, Hess R, Chidambaram V, Kawamoto K, Waller RG, Borbolla D, Del Fiol G, Weir C. Physicians' strategies for using family history data: having the data is not the same as using the data. JAMIA Open. 2020 Oct 1;3(3):378–385. PMID: 34632321; PMCID: PMC7660959.
4. Welch BM, O'Connell N, Schiffman JD. 10 years later: assessing the impact of public health efforts on the collection of family health history. Am J Med Genet A. 2015 Sep;167A(9):2026–2033. PMID: 25939339
5. Polubriaginof F, Tatonetti NP, Vawdrey DK. An Assessment of Family History Information Captured in an Electronic Health Record. AMIA Annu Symp Proc. 2015 Nov 5;2015:2035–2042. PMCID: PMC4765557
6. Ginsburg GS, Wu RR, Orlando LA. Family health history: underused for actionable risk assessment. The Lancet. 2019 Aug 17;394(10198):596–603. PMID: 31395442; PMCID: PMC6822265.
7. Mowery DL, Kawamoto K, Bradshaw R, Kohlmann W, Schiffman JD, Weir C, Borbolla D, Chapman WW, Del Fiol G. Determining Onset for Familial Breast and Colorectal Cancer from Family History Comments in the Electronic Health

Record. AMIA Jt Summits Transl Sci Proc. 2019 May 6;2019:173–181. PMCID: PMC6568127

8. Wood ME, Kadlubek P, Pham TH, Wollins DS, Lu KH, Weitzel JN, Neuss MN, Hughes KS. Quality of Cancer Family History and Referral for Genetic Counseling and Testing Among Oncology Practices: A Pilot Test of Quality Measures As Part of the American Society of Clinical Oncology Quality Oncology Practice Initiative. J Clin Oncol. 2014 Mar 10;32(8):824–829. PMCID: PMC4876350
9. Chavez-Yenter D, Goodman MS, Chen Y, Chu X, Bradshaw RL, Lorenz Chambers R, Chan PA, Daly BM, Flynn M, Gammon A, Hess R, Kessler C, Kohlmann WK, Mann DM, Monahan R, Peel S, Kawamoto K, Del Fiol G, Sigireddi M, Buys SS, Ginsburg O, Kaphingst KA. Association of Disparities in Family History and Family Cancer History in the Electronic Health Record With Sex, Race, Hispanic or Latino Ethnicity, and Language Preference in 2 Large US Health Care Systems. JAMA Network Open. 2022 Oct 4;5(10):e2234574. PMID: 36194411; PMCID: PMC9533178.
10. Weiner S, Amini E, Koeppe E, Resnicow K, Stoffel EM, Griggs JJ. Perceived provider barriers to collecting and documenting a complete family history in a statewide oncology consortium. JCO. Wolters Kluwer; 2021 Oct;39(28_suppl):221–221.
11. Madhavan S, Bullis E, Myers R, Zhou CJ, Cai EM, Sharma A, Bhatia S, Orlando LA, Haga SB. Awareness of family health history in a predominantly young adult population. PLOS ONE. Public Library of Science; 2019 Oct 25;14(10):e0224283. PMID: 31652289; PMCID: PMC6814221.
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv.org. 2018 [cited 2023 Dec 4]. Available from: https://arxiv.org/abs/1810.04805v2
13. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Internet]. arXiv.org. 2019 [cited 2023 Dec 4]. Available from: https://arxiv.org/abs/1907.11692v1
14. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott MBA. Publicly Available Clinical BERT Embeddings [Internet]. arXiv.org. 2019 [cited 2023 Dec 4]. Available from: https://arxiv.org/abs/1904.03323v3
15. Nguyen M, Sedoc J, Taylor CO. Usability, engagement, and report usefulness of chatbot-based family health history data collection: Mixed-methods analysis. Journal of Medical Internet Research (in review). 2023.
16. Brooke J. SUS: A quick and dirty usability scale. Usability Eval Ind. 1995 Nov 30;189.
17. Shen F, Liu S, Fu S, Wang Y, Henry S, Uzuner O, Liu H. Family History Extraction From Synthetic Clinical Narratives Using Natural Language Processing: Overview and Evaluation of a Challenge Data Set and Solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition. JMIR Medical Informatics. 2021 Jan 27;9(1):e24008. PMCID: PMC7875692