

# *FHexchange*: Resources for Family Health History Extraction and Normalization From Consumer Dialog Sources

Michelle H. Nguyen<sup>1\*</sup>, Nidhi Soley<sup>1</sup>, Ayah Zirikly<sup>2,3</sup>, João Sedoc<sup>4</sup>, Casey Overby Taylor<sup>1</sup>

<sup>1</sup> Institute for Computational Medicine, Johns Hopkins University

<sup>2</sup> Center for Language and Speech Processing, Johns Hopkins University

<sup>3</sup> Department of Computer Science, George Washington University

<sup>4</sup> Stern School of Business, New York University

\*Correspondence: {mnguye79} @jhmi.edu

## Abstract

Family health history (FHx) offers insight into a person’s health and disease risk, but it is largely held within free-text clinical formats that require processing for maximal utility of the data. The rapid deployment of ambient AI scribes and conversational agents in clinical settings necessitates evaluation on dynamic patient-clinician and patient-agent dialogs. To address this gap, we introduce two new datasets of patient FHx dialog documents designed to benchmark information extraction and entity linking. Distinct from clinician-entered datasets, patient-reported dialog data has its own semantic and content characteristics, which need to be studied for more patient-centered healthcare. We contribute a publicly available resource called *FHexchange*, with new annotations for family members, clinical observations, related entities, and standardized UMLS CUIs, offering the clinical NLP community a robust evaluation bed for emerging generative AI tools.

## 1 Introduction

Family health history (FHx) is an important health data type that gives insight into genetic, environmental, and behavioral factors that contribute to a person’s disease risk. This information may be collected in a variety of ways, through informational interviews with genetic counselors, primary care providers, and other clinicians, or through asynchronous surveys and chatbots (Wildin et al., 2021). The data collection methods also determine several features of the FHx data, including document and data structure as well as the granularity of data. When FHx information has low granularity (i.e., unstructured symptom descriptions missing standardized concept mappings), it can limit how useful the data are for downstream prediction tasks and clinical decision support (Moon et al., 2022). Thus, to overcome the problem of low granularity, we can extract FHx information in a standardized,

structured format and normalize them to standard vocabularies when possible (Chen et al., 2015).

Much of the collected FHx data still remains within free-text sources; thus, natural language processing (NLP) approaches are necessary to extract this important data (Zhou et al., 2014). Previous work has explored FHx information extraction tasks and relation linking tasks in both synthetic and real-world electronic health record data sources (Shen et al., 2021; Bill et al., 2014; Rybinski et al., 2021; Silva et al., 2020; Goryachev et al., 2008; Zhan et al., 2021), as well as from FHx questionnaires (Shing et al., 2021) using a variety of methods. Methods have included neural language model embeddings with deep learning architectures, rule-based systems, as well as hybrid approaches. These approaches were largely developed on a few public datasets, but more often on proprietary datasets that are not available to the larger clinical NLP research community. Additionally, there are few FHx datasets that provide normalized concepts to support the study of entity linking to clinical codes.

As new AI tools are deployed in the clinic, datasets that reflect new modalities, such as patient-clinician dialogs for ambient AI scribes or conversational AI interactions, are necessary. To support the assessment of FHx extraction for emerging data types, we are contributing the following:

- a newly annotated and normalized resource called *FHexchange* for FHx entity extraction and relation extraction that is publicly available,
- with additional National Library of Medicine’s Unified Medical Language System (UMLS) Concept Unique Identifier (CUI) (U.S. National Library of Medicine) for entity linking benchmarking (that will be accessible for eligible researchers with a UMLS license)

based on two existing dialog corpora: 104 FHx chatbot-user dialogs from (Nguyen et al., 2024), referred to as *FHexchange-KIT*, and 154 FHx-related dialogs from the openly available MTS-Dialog dataset (Ben Abacha et al., 2023), referred to as *FHexchange-MTS*.

Enabled by *FHexchange*, we discuss the following research topics:

- The effect of collection methods on the characteristics and condition coverage of FHx data
- The potential for collection methods to inform what FHx data patients choose to disclose
- The utility of these datasets to support the development of new clinical AI tools.

## 2 Related Work

### 2.1 Existing FHx guidelines and schema

In 2008, the Family Health History Multi-Stakeholder Workgroup of the American Health Information Community convened to formalize a core set of FHx elements. The final core set included person-level attributes (identification, ages, ages/dates at death, cause of death, ethnicity/race, sex, multiple-birth status, adoptive status, relevant genetic/genomic test results), condition-level attributes (disorders list, notation for unknown and/or uncertain histories, and relevant age ranges), and other attributes (consanguinity, identification of biological parents, sensitive data, and integration with other electronic health record data elements) (Feero et al., 2008). This work contributes additional high-quality annotated data for FHx data extraction and concept normalization.

As a part of the Social and Family History - Extraction, Representation, and Evaluation (SFHERE) Project, Chen et al. conducted a multi-source investigation of FHx collection tools and data elements to compile a comprehensive set of relevant FHx entities. They found a wide range of FHx elements represented across three distinct clinical note data sources (MTSamples, University of Pittsburgh Medical Center, and Fletcher Allen Health Care), patient-facing data collection tools (survey of eight tools), and FHx specifications (survey of seven specifications). Through iterative revisions, they designed a highly detailed annotation guideline to direct the manual labeling of clinical notes. The final version of entities and attributes included 13 unique subject categories, nine unique observation categories, four

age/date-related categories, and two categories for other FHx-related information. The annotation resources for the SFHERE Project are publicly available (SFHERE Team, 2017).

Rama et al. (2018) also explored an iterative annotation process for synthetic Norwegian FHx clinical notes. The authors annotated for family members, self (which represents the proband), the condition entries (including diseases, diagnoses, mutations, test results, treatments, and living statuses), and medical events (including acute occurrences). Modifying entities captured included side of the family, age, Negateds, amounts (e.g., *three* sisters, *two* strokes), and temporal expressions (e.g., had a heart attack *two years ago*), which were then related to previously described clinical entities (Rama et al., 2018).

### 2.2 Existing datasets

A summary of datasets that were specifically annotated for FHx mentions can be found in Table 1. Historically, FHx extraction research has heavily relied on clinician-authored or de-identified notes. The MTSamples dataset was used by several groups to explore FHx extraction (Lewis et al., 2011; Bill et al., 2014; Chen et al., 2015); however, only 100 documents are publicly accessible as of 2026 from the SFHERE Project (2014). The 2019 n2c2/OHNLP and BioCreative/OHNLP challenge datasets (Shen et al., 2021; Rastegar-Mojarad et al., 2018) also provided foundational benchmarks for extracting entities and relations, from highly detailed genetics-specialty-style notes. Wang et al. reannotated the datasets to correct errors in the original version as well as to enhance the data by normalizing observations to SNOMED-CT codes and creating a FHx lexical resource (Wang et al., 2023). Both corpora are only accessible through a data use agreement (DUA). Rama et al. continued their work and released the dataset of 477 cardiac-related FHx sentences (in Norwegian) in 2021 (Brekke et al., 2021), which is freely accessible.

Recent work has explored consumer-reported and patient-involved datasets. Azab et al. (2019) created a FHx dataset of 228 patient-facing questionnaires from genetic counseling sessions and crowd-sourced participants. This dataset was very rich and information-dense, containing over 20,000 FHx relations; however, the dataset is no longer accessible. Saley et al. (2024) released MediTOD in 2024 based off a publicly available set of patient-

Dataset	Type/Source	Language	FHx annotation types	Disease focus	Size of FHx-related content	FHx annotation access*
<b>Clinician-authored sources</b>						
MTSamples (MT-Samples)	Synthetic clinical note	English	Clinical entity identification (family members, observations, living status, ages), entity properties (such as side of family, Negated), and relation extraction.	Variety	491-3940 documents	Freely accessible for 100 documents (SFHERE Project, 2014)
2019 n2c2/OHNLP FH (Shen et al., 2021); BioCreative/OHNLP 2018 (Rastegar-Mojarad et al., 2018)	Synthetic clinical note	English	Clinical entity identification (family members, observations, living status, ages), entity properties (side of family, Negated), and relation extraction.	Variety/ Genetics	216 documents	DUA required (99 documents, training set)
Brekke et al. (2021)	Synthetic clinical note	Norwegian	Clinical entity identification (family members, event, conditions), entity properties (ages, side of family, Negated, amount, temporal), and relation extraction	Cardiology	477 sentences	Freely accessible
2019 n2c2/OHNLP FH; BioCreative/OHNLP 2018 re-annotated (Wang et al., 2023)	Synthetic clinical note	English	Clinical entity identification (family members, observations, living status, ages), entity properties (side of family, Negated), relation extraction, <b>normalized concepts (SNOMED CT)</b> .	Variety/ Genetics	216 documents	DUA required
<b>Consumer-reported sources</b>						
Azab et al. (2019)	Real patient-facing questionnaire	English	Clinical entity identification (family members, illnesses, cancer, cancer type), entity properties (living status, sex, ages), and relation extraction.	Cancer	228 questionnaires	No longer accessible
MediTOD (Fareez et al., 2022; Saley et al., 2024)	Synthetic patient-clinician dialog	English	Clinical entity identification (family members, observations), entity properties (status, Negated), relation extraction, <b>normalized concepts (UMLS CUI)</b> .	Respiratory	214 documents	Baseline annotations (freely accessible), Normalized annotations (UMLS license required)
<i>FHexchange-MTS</i> , from MTS-DIALOG subset (Ben Abacha et al., 2023)	Synthetic patient-clinician dialog	English	Clinical entity identification (family members, observations, living status, ages), entity properties (e.g., side of family, Negated), relation extraction, <b>normalized concepts (UMLS CUI)</b>	Variety	154 documents	Baseline annotations (freely accessible), Normalized annotations (UMLS license required)
<i>FHexchange-KIT</i> , from KIT-user dialogs (Nguyen et al., 2024)	Real user-chatbot interaction transcript	English	Clinical entity identification (family members, observations), relation extraction, <b>normalized concepts (UMLS CUI)</b>	Variety	104 documents	Baseline annotations (freely accessible), Normalized annotations (UMLS license required)

Table 1: Snapshot of released datasets by source type with FHx entity, relation annotations, and **normalized concepts**, with our contributions **highlighted**, \*access as of 2026

clinician dialogs from Fareez et al. (2022). The authors labeled dialogs for conversational attributes, such as the intent of the dialog turn, as well as topic areas, termed by the authors as *slots*, which could be information regarding symptoms, collection of patient medical history, or collection of family med-

ical history, which was of particular interest for the present work. There are 214 dialogs containing FHx mentions, and the authors primarily labeled for family member, condition, status and Negated, leaving other FHx details in a semi-structured text field. To support additional evaluation of clinical

```

Chat Transcript (2/14)
14 [2023-01-10 07:21:00 (American,English)] user: [{"type": "Action", "id": "question_213", "question": "question_213", "value": 0, "text": "Cancer conditions"}]
15 [2023-01-10 07:21:00 (American,English)] kit: [{"type": "Action", "id": "question_213", "question": "question_213", "value": 0, "text": "Cancer conditions"}]
16 [2023-01-10 07:21:01 (American,English)] kit: [{"type": "Form", "data": [{"type": "Form", "instruction": "", "display-type": "list", "id": "question_4", "questions": [{"content-var-name": "q", "content": [{"text": "Spouse/child of your family members have ever been diagnosed with a type of cancer?<br/>Select all that apply from the following list.<br/>"}], "choices": [{"text": "Mother", "value": 0}, {"text": "Father", "value": 1}, {"text": "Sibling", "value": 2}, {"text": "Daughter", "value": 3}, {"text": "Son", "value": 4}, {"text": "Grandparent", "value": 5}], "text": "Spouse/child of your family members have ever been diagnosed with a type of cancer?<br/>Select all that apply from the following list.<br/>"}, {"text": "Multiple choice", "id": "question_4", "topic-id": 4, "required": true, "heading": "Cancer conditions"}]}]}]
17 [2023-01-10 07:21:15 (American,English)] user: [{"type": "Action", "id": "question_4", "question": "question_4", "value": 5, "text": "Grandparent"}]
18 [2023-01-10 07:21:15 (American,English)] kit: "Thanks for your input."
19 [2023-01-10 07:21:16 (American,English)] kit: [{"type": "Action", "id": "question_4", "question": "question_4", "value": 5, "text": "Grandparent"}]
20 [2023-01-10 07:21:16 (American,English)] kit: [{"type": "Form", "data": [{"type": "Form", "instruction": "", "display-type": "list", "id": "question_198", "questions": [{"content-var-name": "q", "content": [{"text": "What type of cancer has your grandparent had? Select all that apply. Ask for help if you are unsure what one of these items means.<br/>"}], "min-input-len": 1, "choices": [{"text": "Bladder cancer", "value": 0}, {"text": "Breast cancer", "value": 1}, {"text": "Colon/rectal cancer", "value": 2}, {"text": "Endometrial cancer", "value": 3}, {"text": "Kidney cancer", "value": 4}, {"text": "Lung cancer", "value": 5}, {"text": "Pancreatic cancer", "value": 6}, {"text": "Skin cancer", "value": 7}, {"text": "Stomach cancer", "value": 8}, {"text": "Other", "value": 9}], "text": "What type of cancer has your grandparent had? Select all that apply. Ask for help if you are unsure what one of these items means."}, {"text": "Multiple choice", "id": "question_198", "topic-id": 198, "required": true, "heading": "Grandparent - cancer"}]}]}]
21 [2023-01-10 07:21:21 (American,English)] user: [{"type": "Action", "id": "question_198", "question": "question_198", "value": 1, "text": "Breast cancer"}]
22 [2023-01-10 07:21:21 (American,English)] kit: "OK!"
23 [2023-01-10 07:21:22 (American,English)] kit: [{"type": "Action", "id": "question_198", "question": "question_198", "value": 1, "text": "Breast cancer"}]

```

(a) KIT Transcript from a user interaction. In the red box, we can identify that the user has indicated that their family has a history of cancer conditions, specifically that their grandparent had breast cancer.

```

Doctor: Tell me about your family. Did anyone have any medical issues?
Patient: My father is alive. He is sixty nine right now. He did have some strokes in the past.
My mom on the other hand died at the age of sixty two of many kinds of heart diseases.
Doctor: What about your grandparents?
Patient: My maternal grandfather he also had heart issues and he died of a heart attack.
Both he and my maternal grandma had some heart diseases. Their heart use to beat pretty fast.
FamilyMember LivingStatus SideoftheFamily Observation AgeatDeath Age

```

(b) MTS-Dialog sample patient-doctor conversation. The annotated entities are highlighted.

Figure 1: Examples of KIT and MTS-Dialog data

concepts, Saley et al. (2024) mapped observations to UMLS codes using QuickUMLS (Soldaini and Goharian).

We contribute *FHexchange*, a new dataset consisting of two additional consumer-reported sources derived from: (1) the MTS-Dialog dataset (Ben Abacha et al., 2023) and (2) KIT-user dialogs (Nguyen et al., 2024). The full MTS-Dialog dataset consists of 1701 synthetic patient-doctor conversation documents and summarized clinical notes for these dialogs. Because the original dataset was designed to support clinical note summarization, the authors did not annotate for specific elements such as FHx. To construct *FHexchange-MTS*, we selected a set of 154 MTS-Dialog dialogs that contained FHx mentions for further FHx annotation, a sample dialog can be seen in Figure 1b. For the second subdataset, *FHexchange-KIT*, we annotated 104 KIT-user dialogs from a crowdsourced study to assess chatbot-based data collection from a chatbot named KIT (Nguyen et al., 2024). The questionnaire was adapted from the NIH *AllofUs* family history survey and covered intergenerational FHx for 11 pre-defined condition categories (The All of Us Research Program Investigators, 2019). Figure 1a shows an example of an *FHexchange-KIT* excerpt.

### 3 FHexchange dataset

#### 3.1 Annotation

The annotation guideline was created by referencing existing schema (Chen et al., 2015) for a subset of terms and further development with clinical experts (genetic counselors). The annotation guideline was iteratively refined by annotating a sample set of clinical notes with guidance from two genetic

counselors, each with over ten years of clinical experience.

The *FHexchange-MTS* data was annotated for the following concepts: *FamilyMember*, *Observation*, *SideoftheFamily*, *LivingStatus*, *Age*, *AgeofOnset*, *AgeofDeath*, *CauseofDeath*, and *Negated*, following a schema further described in Appendix A. The *FHexchange-KIT* data consisted of user-entered health conditions for a specified set of *FamilyMembers*: Father, Mother, Sibling, Grandparent, Son, and Daughter. Because of this, the annotations were limited to *FamilyMember* and *Observation*. Two Masters-level individuals (both native English speakers) trained in Biomedical Engineering completed the clinical entity annotations. Both annotators had overlap on the full corpus.

UMLS concept normalization was first conducted using the python library, scispacy’s UMLS entity linker, with a link score (calculated as a nearest neighbor search percentage between the candidate text and the UMLS CUIs) (Neumann et al., 2019). Only matches with a link score of over 0.9 were retained. One human annotator also normalized all concepts for both *FHexchange-MTS* and *FHexchange-KIT* clinical observations for comparison with automated code assignments.

#### 3.2 FHexchange dataset summary

We report interrater agreement using Cohen’s kappa for the presence of *Observations* and *FamilyMembers*. We also report percent agreement for both datasets. For the *FHexchange-KIT* subdataset, we report percent agreement on a document-level as the annotations were limited to a fixed set of *FamilyMembers*. Percent agreement was calculated as the proportion of overlapping observations between annotators across six predefined *Family-*

*Members*, while full agreement strictly required perfectly identical observation sets for every family member within the transcript.

To report coverage of FHx conditions within the dataset, we performed further linking of UMLS concepts to Semantic Types and ICD-10 Body Systems using UMLS API calls with UMLS version 2025AB. We also reported the distribution of UMLS semantic types and ICD-10 body systems present. All computation was performed using python version 3.11.

### 3.3 Example baselines with zero-shot and few-shot large language model structured output approaches

We demonstrate the value of the dataset by providing an example baseline for structured FHx information extraction, relation extraction, and entity linking using the following pre-trained language models with structured outputs: Microsoft Foundry Azure OpenAI zero-shot GPT-4o (Version 2024-11-20) (Microsoft, a), GPT-5-mini (Version 2025-08-07) (Microsoft, b), and Microsoft Azure Llama-3.3-70B-Instruct (Version 9) (Microsoft, c). We aimed to select proprietary models (flagship model versus lightweight reasoning model) and an open-weight model. Although we did not include a clinical model baseline, we included an open-weight model to represent a solution that could be locally deployed, considering data privacy and compliance with healthcare regulations. We created a structured Pydantic (Colvin et al., 2026) schema to prompt the model to reason through FHx extraction (with a *reasoning* field as the first entry in the schema) of the relevant FHx annotated entities where available (*FamilyMember*, *Observation*, *AgeofOnset*, *SideoftheFamily*, *LivingStatus*, *Age*, *AgeofDeath*, *CauseofDeath*, *CUI*, and *Negated*). Model, prompt details, and schema description can be found in the Appendix, and additional details are available on the dataset Github repository.

To handle the structural differences between *FHexchange-KIT* and *FHexchange-MTS*, the schema utilized optional typing, allowing the models to output null for variables unsupported by the text without breaking the JSON generation. Additionally, prompt constraints were applied to standardize clinical concepts, such as negated histories (e.g., forcing the extraction of the base condition with a *Negated: true* boolean flag) and restricting categorical modifiers to predefined values (such as Maternal, Paternal, or Unknown, for *SideoftheFam-*

*ily*).

After normalizing the model output, evaluation was performed by reporting precision, recall, and F1 for all entities extracted and all relations extracted. Matches for *Observations* and *FamilyMembers* were evaluated strictly based off perfect token-level matches compared to the gold standard annotations.

Metric	Value
<i>FHexchange-MTS</i>	
Binary Presence ( $\kappa$ )	0.827
Strict match ( $\kappa$ )	0.300
Lenient match ( $\kappa$ )	0.519
<i>FHexchange-KIT</i>	
Binary Presence ( $\kappa$ )	0.990
Perfect match (%)	63.5%
Partial match (%)	36.5%

Table 2: Inter-annotator agreement across datasets. For *FHexchange-MTS*, we report phrase-level  $\kappa$  representing strict and lenient text span matches. For *FHexchange-KIT*, we report  $\kappa$  for the presence of *FamilyMember* and *Observation* fields and raw document-level percent agreement for exact sets of conditions. Additional agreement details for *Observation-FamilyMember* relations can be seen in Appendix Table 9.

Concept	<i>FHexchange-MTS</i>	<i>FHexchange-KIT</i>
<i>Dataset Statistics</i>		
Documents	154	104
Dialog turns	856	5156
<i>Annotation Counts</i>		
FamilyMember	265	249
Observation	488	720
SideoftheFamily	24	–
LivingStatus	283	–
Age	20	–
AgeofOnset	12	–
AgeofDeath	59	–
CauseofDeath	70	–
CUI	424	679
Negated	112	–
<i>Relations (Total)</i>	1280	1399

Table 3: Annotation statistics for *FHexchange-MTS* and *FHexchange-KIT*

## 4 Results

Appendix Table 2 details the results of annotator agreement for the two datasets. For *FHexchange-MTS*, Cohen’s kappa for the overlap regarding the presence of the entities: *FamilyMember* type, *LivingStatus* type, or *SideoftheFamily* type and *Observation*, was high (0.827, indicating a near perfect

Rk	Condition	N	Sample Expressions	Category	FHex-KIT	FHex-MTS
<i>FHexchange-MTS</i>						
1	Diabetes mellitus	46	<i>diabetes, has sugar</i>	Unique CUIs	114	121
2	Heart diseases	43	<i>heart disease, heart issues, heart problems</i>	<i>Semantic Types (FHex-KIT: 8, FHex-MTS: 13)</i>		
3	Hypertension	26	<i>high blood pressure, hypertension</i>	Disease or Syndrome	68	54
4	Cerebrovascular accident	23	<i>stroke, CVA</i>	Neoplastic Process	22	30
5	Healthy	20	<i>healthy, no conditions</i>	Mental/Behavioral Dysfunction	11	10
<i>FHexchange-KIT</i>						
1	High blood pressure - hypertension	53	-	Pathologic Function	6	6
2	Depression	37	-	Finding	3	7
3	Heart attack	28	-	Sign or Symptom	2	4
4	High cholesterol	26	-	Anatomical Abnormality	1	3
5	Anxiety reaction/panic disorder	21	-	Congenital Abnormality	1	-
<i>ICD-10 Body Systems (FHex-KIT: 12, FHex-MTS-Dialog: 11)</i>						
No Direct ICD-10 Mapping					26	48
Neoplasms (Oncology)					16	18
Other/General					12	7
Nervous System					12	12
Digestive System					11	3
Mental and Behavioral					9	6
Endocrine/Nutritional/Metabolic					8	6
Circulatory System					7	12
Musculoskeletal System					7	4
Respiratory System					3	4
Genitourinary System					2	2
Blood/Blood-forming Organs					1	-

Table 4: Top 5 conditions by subdataset with sample patient expressions. Top conditions for *FHexchange-KIT* did not contain any variance, so no sample expressions are listed.

match) (Landis and Koch, 1977). When comparing a strict phrase-level match, where tokens must perfectly overlap, the kappa value was low (0.300); however, when allowing for fuzzy matches for a lenient phrase-level match, the agreement improves between annotators (0.519). Appendix B provides examples of errors between annotators. The performance in the results for the strict match were lower due to differences in the boundaries annotators notated (Table 9 provides additional details regarding categories of interannotator disagreement). For example, for age at death, one annotator specified *AgeofDeath* was "40s", but another specified this was "died in 40s." For this example, we removed all modifying text related to *AgeofDeath*, to only retain "40s." For *FHexchange-KIT*, there was high agreement for the presence of *FamilyMember* and *Observation* with a Cohen’s  $\kappa$  of 0.990. For document-level agreement, 63.5% of the documents were perfect matches between annotators.

Table 3 contains the annotation statistics across both datasets. *FHexchange-MTS* contains 856 dialog turns, and *FHexchange-KIT* contains 5156 dialog turns.

Table 4 presents the top conditions and frequencies after mapping *Observations* to CUIs across both datasets. Sample expressions are provided for *FHexchange-MTS* to reflect the variations in patient descriptors for diseases. The two subdatasets both have hypertension in the top five conditions.

When observing higher-order categories of conditions, we identified eight unique Semantic Types of mapped CUIs from the *FHexchange-KIT* subdataset and 13 Semantic Types for *FHexchange-MTS*, as seen in Table 5. We found 12 unique ICD-10 Body systems for *FHexchange-KIT* and 11

Table 5: UMLS semantic types and ICD-10 body system distributions across *FHexchange-KIT* (*FHex-KIT*) and *FHexchange-MTS* (*FHex-MTS*)

Body systems for *FHexchange-MTS*.

In Table 6, we report baseline performance of the two proprietary pre-trained language models, Azure OpenAI GPT-4o and GPT-5-mini with zero-shot and few-shot in-context learning approaches, to extract family history information and to perform relation extraction and UMLS entity linking. The full experimental results with performance from all three models (including Azure Llama 3.3 70B Instruct) can be seen in Appendix Section D. Zero-shot prompting performance on extracting *FamilyMembers* for both subdatasets is high for both models at baseline (GPT-4o: F1 = 0.994 for *FHexchange-KIT*, F1 = 0.851 for *FHexchange-MTS*; GPT-5-mini: F1 = 0.972 for *FHexchange-KIT*, F1 = 0.774 for *FHexchange-MTS*). However, performance declined for both with few-shot prompting for the *FHexchange-KIT*

		FM	Obs	Rel	CUI	Neg	Living	Side
<i>FHexchange-KIT</i>								
GPT-4o	ZS	.994 (.99/1.0)	.578 (.53/.64)	.632 (.66/.61)	.354 (.36/.35)	.573 (.52/.64)	—	—
	FS	.986 (.98/.99)	.638 (.58/.71)	.702 (.71/.69)	.373 (.37/.38)	.039 (.08/.03)	—	—
GPT-5-mini	ZS	.972 (.97/.97)	.726 (.66/.81)	.824 (.85/.80)	.115 (.48/.07)	.714 (.65/.79)	—	—
	FS	.891 (.97/.82)	.649 (.63/.67)	.711 (.79/.65)	.255 (.44/.18)	.000	—	—
<i>FHexchange-MTS</i>								
GPT-4o	ZS	.851 (.97/.76)	.546 (.64/.48)	.528 (.61/.47)	.597 (.61/.59)	.490 (.55/.44)	.681 (.66/.70)	.400 (.67/.29)
	FS	.873 (.97/.79)	.692 (.79/.62)	.695 (.74/.66)	.564 (.56/.57)	.269 (.50/.18)	.641 (.59/.70)	.133 (1.0/.07)
GPT-5-mini	ZS	.774 (.79/.76)	.674 (.75/.61)	.625 (.67/.59)	.234 (.80/.14)	.612 (.65/.58)	.740 (.80/.69)	.429 (.60/.33)
	FS	.881 (.94/.83)	.670 (.74/.61)	.658 (.68/.64)	.553 (.63/.49)	.235 (.46/.16)	.705 (.69/.72)	.000

Table 6: Baseline performance of proprietary models. Metrics: F1 (Precision/Recall). FM = *FamilyMember*, Obs. = *Observation*, Rel = FM-Obs Relation, Neg. = *Negated*, Living = *LivingStatus*, Side = *SideoftheFamily*. ZS = Zero-Shot, FS = Few-Shot.

subdataset, but increased for the *FHexchange-MTS* subdataset, with GPT-5-mini improving performance over GPT-4o with the provided examples. Overall, GPT-5-mini had greater zero-shot performance to extract *Observations*, which translated to stronger performance to relate *FamilyMembers* to *Observations*. Across both models, performance to extract age-related concepts was poor for *FHexchange-MTS*, missing all mentions of *Age*, *AgeofOnset*, and *AgeofDeath*. UMLS entity linking (CUI normalization) remained a difficult task across both subdatasets (best F1 = 0.373 for *FHexchange-KIT* with few-shot GPT-4o and F1 = 0.597 for *FHexchange-MTS* with zero-shot GPT-4o). In Appendix Table 13, we show three examples of differences between the models in extracting CUIs. GPT-5-mini was more conservative in assigning CUIs to *Observations* than GPT-4o. Few-shot prompting, where example CUI mappings were provided, is hypothesized to have increased the performance of GPT-5-mini to assign CUIs (0.115 to 0.255 for *FHexchange-KIT* and 0.234 to 0.553 for *FHexchange-MTS*).

As detailed in Appendix Tables 11 for *FHexchange-KIT* and 12 for *FHexchange-MTS*, the performance of Llama 3.3 was poor for all entities in comparison to the proprietary models. This was due to structural output errors in which the model did not correctly adhere to the JSON schema, despite correctly identifying entities within the reasoning field; however, few-shot prompting did improve the performance. This is a limitation, which may be addressed through additional prompt examples, fine-tuning, or parsing and normalization, and is an area for future work.

## 5 Discussion

### The effect of collection methods on the characteristics and condition coverage of FHx data

To enable more patient-centered care, there is a need for datasets derived from patient input. Patients report health data that is unique from their clinicians, providing additional context for a more holistic health profile (Castro et al., 2025). Data collection methods (surveys versus in-person medical history interviews) inform the characteristics and quality of the resulting dataset (Chen et al., 2015). Data from surveys with multiple choice options normalize the range of conditions and family members collected, due to the limitation of choices, whereas open-ended dialogs can lead to responses with more variance. Together, *FHexchange-MTS* and *FHexchange-KIT* provide a complementary view of modern patient-derived data for FHx, originating from their collection methods.

*FHexchange-MTS* is based on open-ended, free-form conversations about FHx. Not all dialogs contain specific prompting by doctors to patients, and in general, patients are tasked with spontaneously recalling relevant FHx. Because of this data collection method, we identify that *FHexchange-MTS* contains more colloquial vernacular, such as "sugar" for diabetes and "alcoholism" for alcohol use disorder. We provide mappings to the relevant CUI codes to standardize this language to computable clinical concepts. Additionally, the breadth of reported information was greater for *FHexchange-MTS* in comparison to *FHexchange-KIT* (13 UMLS Semantic Types captured versus eight Semantic Types).

In contrast, KIT users were prompted to recall their FHx for specific condition types and with

multiple choice options during their dialogs, and therefore, users relied more on option selection than free-text inputs. Because of this data collection method, the majority of responses within *FHexchange-KIT* were mapped to distinct options, leaving little variation in how reported conditions were expressed. When cross-mapping from UMLS codes to the ICD-10 clinical hierarchy, 39% of the unique concepts extracted from the open-ended *FHexchange-MTS* lacked a direct ICD-10 equivalent, compared to only 22% in the structured *FHexchange-KIT* survey data. This difference reinforces that open-ended dialogs yield a significantly wider variance of informal patient expressions, and structured dialogs may enable cleaner extraction into strict diagnostic hierarchies and computable formats to support automated risk assessments.

**The potential for collection methods to inform what FHx data patients choose to disclose** The open-ended dialogs in *FHexchange-MTS* show high enrichment for Circulatory System and Neoplasm-related disorders (12 and 18 unique codes respectively). The structured KIT dialogs show deeper representation across chronic condition systems like Digestive System (11 unique concepts, as compared to *FHexchange-MTS*'s three unique Digestive concepts). This may also suggest that asynchronous, comprehensive, structured methods of collecting FHx may be more effective to elicit both a broader (capturing a wider range of condition categories) and more specific (capturing finer grained condition language) array of FHx data than possible within limited conversations in a non-specialist setting. Asynchronous surveys conducted through chatbots may also lead to a higher willingness to share sensitive information when compared to face-to-face interactions with healthcare providers (Lucas et al., 2014). Supporting this notion, we found that two of the top five conditions represented within *FHexchange-KIT* were mental health related conditions, which patients may experience a stigma to report in traditional doctor-patient interactions (Nadarzynski et al., 2019). However, the populations and clinical settings represented in these datasets may also be distinct from one another, so future work should explore FHx data collection methods across the same populations and clinical contexts.

**The utility of *FHexchange* to support the development of new clinical AI tools** In the future, clinical AI tools will capture FHx across multiple

modalities, from ambient scribes transcribing and documenting natural patient-doctor interactions to more flexible conversational agents (text and audio) providing medical history taking interviews (Hindelang et al., 2024). We contribute *FHexchange*, a unified FHx dialog dataset of both structured chatbot-user dialogs as well as unstructured human-human dialogs. The baseline evaluations of proprietary and open-weight models support the value of this corpus. While the proprietary language models accurately extracted basic family members from structured inputs, they did not extract temporal modifiers (ages) and other condition modifiers (*CauseofDeath*). Additionally, the UMLS CUI entity linking performance suggests that the field still needs robust, standardized mapping workflows to achieve computable clinical concepts for FHx. Error analysis revealed that the model translated colloquial patient expressions (e.g., "kidney cancer" and "fast heart beat") into formal medical terminology ("renal cancer" and "tachycardia"). In future work, we will consider these conceptually equivalent extractions as true positives; however, this highlights a tendency for the model to perform clinical normalization and not just exact extraction of the text. Additionally, our results suggest that reasoning models like GPT-5-mini may be conservative in linking clinical observations to standard terminologies that reduce the linking of both incorrect and correct mappings. These findings motivate the work to introduce additional domain knowledge, through the integration of ontologies or knowledge graphs, to ground the mapping of family history mentions to standard terminologies in future studies (Gao et al., 2023, 2025; Pan et al., 2024). Furthermore, future work could explore fine-tuning open-weight models to improve performance, as even few-shot prompts boosted Llama 3.3's ability to correctly extract FHx information (Park et al., 2023). *FHexchange* provides an evaluation set to assess emerging clinical AI tools across data collection modalities.

## 6 Conclusion

We are contributing *FHexchange*, consisting of annotated FHx resources for entity identification, relation extraction, and concept normalization from two distinct consumer-facing sources. The base data (for entity identification and relation extraction) will be available publicly for the clinical NLP research community from this data repository: [www.](http://www.)

[github.com/tirilab/FHexchange](https://github.com/tirilab/FHexchange). By combining structured chatbot dialogs with unstructured human-to-human dialogs, we provide a comprehensive FHx dataset designed to benchmark the performance of emerging clinical AI systems to extract and normalize FHx data across multiple conversational modalities.

## Limitations

First, the quality of patient-reported health information is characterized by patient recall. We did not annotate entities for patients' certainty of recall, but this consideration has been explored in previous annotation schema (Chen et al., 2015). Second, we did not include additional inter-annotator agreement metrics, such as an annotator confusion matrix. Third, the evaluation of the baseline is currently on strict token-level matches of *Observations*; therefore, this is currently underestimating the performance of the models. Fourth, the models were prompted for structured output by providing a structured Pydantic schema for extraction. This approach has shown to decrease performance and model reasoning (Tam et al., 2024), and we did not process Llama 3.3's structured outputs beyond the processing performed for the OpenAI models in order to extract the usable information. Fifth, baseline results did not include a clinical baseline. We only performed one model run for each document for each model, and the temperature of the models was not set to 0. Therefore, further testing of the consistency of the results is needed, as well as exploration of other models for this task. Sixth, we did not specifically annotate for strength of observations (e.g., indications that many family members experience a similar condition type or the perceived severity of certain observations). Seventh, for these datasets, we focus on English language dialogs only. Future work to support other languages is necessary. Lastly, UMLS CUIs do not capture all commonly used patient and/or consumer language for conditions. Previous work has been done to add consumer health vocabulary that was not explored in this work (Pham et al., 2019; Amith et al., 2019).

## Ethical Considerations

All chatbot transcripts were anonymized to remove any identifying information and timestamps. We utilize the UMLS Metathesaurus to standardize medical observations via sciSpacy's entity linking module. We accessed the Metathesaurus (ver-

sion 2022AB in scispacy after 2023 and version 2025AB through the API) through an individual license obtained for this work. Although the base annotations of FHx will be freely available for the research community, the UMLS CUI linkages will be provided through protected access that ensures each researcher abides by UMLS licensing requirements.

## Acknowledgments

We would like to thank the genetic counselors and annotators who donated their time to make this work possible.

## References

- Muhammad Amith, Kirk Roberts, Cui Tao, Licong Cui, and Hua Xu. 2019. *Ontology of Consumer Health Vocabulary: providing a formal and interoperable semantic resource for linking lay language and medical terminology*. *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, 2019:1177–1178.
- Mahmoud Azab, Stephane Dadian, Vivi Nastase, Larry An, and Rada Mihalcea. 2019. *Towards Extracting Medical Family History from Natural Language Interactions: A New Dataset and Baselines*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1255–1260, Hong Kong, China. Association for Computational Linguistics.
- Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. *An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters*. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.
- Robert Bill, Serguei Pakhomov, Elizabeth S. Chen, Tamara J. Winden, Elizabeth W. Carter, and Genevieve B. Melton. 2014. *Automated Extraction of Family History Information from Clinical Notes*. *AMIA Annual Symposium Proceedings*, 2014:1709–1717.
- Pål H. Brekke, Taraka Rama, Ildikó Pilán, Øystein Nytrø, and Lilja Øvrelid. 2021. *Synthetic data for annotation and extraction of family history information from clinical text*. *Journal of Biomedical Semantics*, 12(1):11.
- Victor M. Castro, Vivian S. Gainer, Danielle M. Crookes, Shawn N. Murphy, and Justin Manjourides. 2025. *Comparing patient-reported symptoms and structured clinician documentation in electronic*

- [health records](#). *Journal of the American Medical Informatics Association: JAMIA*, 32(9):1454–1461.
- Elizabeth S Chen, Elizabeth W Carter, Tamara J Winden, Indra Neil Sarkar, Yan Wang, and Genevieve B Melton. 2015. [Multi-source development of an integrated model for family health history](#). *Journal of the American Medical Informatics Association : JAMIA*, 22(e1):e67–e80.
- Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, Alex Hall, and Victorien Plot. 2026. [Pydantic Validation](#).
- Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. [A dataset of simulated patient-physician medical interviews with a focus on respiratory cases](#). *Scientific Data*, 9(1):313.
- W. Gregory Feero, Mary Beth Bigley, and Kristin M. Brinner. 2008. [New Standards and Enhanced Utility for Family Health History Information in the Electronic Health Record: An Update from the American Health Information Community’s Family Health History Multi-Stakeholder Workgroup](#). *Journal of the American Medical Informatics Association : JAMIA*, 15(6):723–728.
- Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. 2023. [DR.BENCH: Diagnostic Reasoning Benchmark for Clinical Natural Language Processing](#). *Journal of biomedical informatics*, 138:104286.
- Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. [Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study](#). *JMIR AI*, 4:e58670.
- Sergey Goryachev, Hyeoneui Kim, and Qing Zeng-Treitler. 2008. [Identification and Extraction of Family History Information from Clinical Reports](#). *AMIA Annual Symposium Proceedings*, 2008:247–251.
- Michael Hindelang, Sebastian Sitaru, and Alexander Zink. 2024. [Transforming Health Care Through Chatbots for Medical History-Taking and Future Directions: Comprehensive Systematic Review](#). *JMIR Medical Informatics*, 12(1):e56628.
- J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Neal Lewis, Daniel Gruhl, and Hui Yang. 2011. [Dependency Parsing for Extracting Family History](#). In *2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 237–242.
- Neal Lewis, Daniel Gruhl, Hui Yang, San Francisco, San Jose, and San Francisco. [Extracting Family History Diagnoses From Clinical Texts](#).
- Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. [It’s only a computer: Virtual humans increase willingness to disclose](#). *Computers in Human Behavior*, 37:94–100.
- Microsoft. a. [gpt-4o](#).
- Microsoft. b. [gpt-5-mini](#).
- Microsoft. c. [Llama-3-3-70B-Instruct](#).
- Sungrim Moon, Liwei Wang, Xuan Chen, Nan Wang, Sheila M. Manemann, Nicholas B. Larson, Suzette J. Bielinski, and Hongfang Liu. 2022. [Bridging the Granularity Gap in Family History Information Extracted from Clinical Narratives](#). *AMIA ... Annual Symposium proceedings. AMIA Symposium, 2022:795–804*.
- MTSamples. [Medical Transcription Sample Reports & Examples | MTSamples](#).
- Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. 2019. [Acceptability of artificial intelligence \(AI\)-led chatbot services in healthcare: A mixed-methods study](#). *Digital Health*, 5.
- Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. [ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing](#). *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327. Conference Name: *Proceedings of the 18th BioNLP Workshop and Shared Task*.
- Michelle Hoang Nguyen, João Sedoc, and Casey Overby Taylor. 2024. [Usability, Engagement, and Report Usefulness of Chatbot-Based Family Health History Data Collection: Mixed Methods Analysis](#). *Journal of Medical Internet Research*, 26(1):e55164. Company: *Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research*.
- Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. [Unifying Large Language Models and Knowledge Graphs: A Roadmap](#). *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.
- Hyeryun Park, Jiye Son, Jeongwon Min, and Jinwook Choi. 2023. [Selective UMLS knowledge infusion for biomedical question answering](#). *Scientific Reports*, 13(1):14214.

- Quynh Pham, Gary Graham, Carme Carrion, Plinio P. Morita, Emily Seto, Jennifer N. Stinson, and Joseph A. Cafazzo. 2019. [A Library of Analytic Indicators to Evaluate Effective Engagement with Consumer mHealth Apps for Chronic Conditions: Scoping Review](#). *JMIR mHealth and uHealth*, 7(1):e11941. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth.
- Taraka Rama, Pål Brekke, Øystein Nytrø, and Lilja Øvrelid. 2018. [Iterative development of family history annotation guidelines using a synthetic corpus of clinical text](#). In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 111–121, Brussels, Belgium. Association for Computational Linguistics.
- Majid Rastegar-Mojarad, Sijia Liu, Yanshan Wang, Naveed Afzal, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu. 2018. [BioCreative/OHNL Challenge 2018](#). In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '18*, page 575, New York, NY, USA. Association for Computing Machinery.
- Maciej Rybinski, Xiang Dai, Sonit Singh, Sarvnaz Karimi, and Anthony Nguyen. 2021. [Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis](#). *JMIR Medical Informatics*, 9(4):e24020.
- Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam . 2024. [MediTOD: An English Dialogue Dataset for Medical History Taking with Comprehensive Annotations](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.
- SFHERE Project. 2014. [Sample sentences and annotations for family history](#).
- SFHERE Team. 2017. [Digital Library for Social and Family History-Extraction-Representation, and Evaluation Project](#).
- Feichen Shen, Sijia Liu, Sunyang Fu, Yanshan Wang, Sam Henry, Ozlem Uzuner, and Hongfang Liu. 2021. [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 \(OHNL\) Competition](#). *JMIR medical informatics*, 9(1):e24008.
- Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. 2021. [Towards Clinical Encounter Summarization: Learning to Compose Discharge Summaries from Prior Notes](#). *arXiv preprint*. ArXiv:2104.13498 [cs].
- João Figueira Silva, João Rafael Almeida, and Sérgio Matos. 2020. [Extraction of Family History Information From Clinical Notes: Deep Learning and Heuristics Approach](#). *JMIR Medical Informatics*, 8(12):e22898.
- Luca Soldaini and Nazli Goharian. [QuickUMLS: a fast, unsupervised approach for medical concept extraction](#).
- Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. [Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models](#). *arXiv preprint*. ArXiv:2408.02442 [cs].
- The All of Us Research Program Investigators. 2019. [The “All of Us” Research Program](#). *The New England journal of medicine*, 381(7):668–676.
- U.S. National Library of Medicine. [Unified Medical Language System \(UMLS\)](#).
- Liwei Wang, Huan He, Andrew Wen, Sungrim Moon, Sunyang Fu, Kevin J. Peterson, Xuguang Ai, Sijia Liu, Ramakanth Kavuluru, and Hongfang Liu. 2023. [Acquisition of a Lexicon for Family History Information: Bidirectional Encoder Representations From Transformers–Assisted Sublanguage Analysis](#). *JMIR Medical Informatics*, 11(1):e48072.
- Robert S. Wildin, Donna J. Messersmith, and Elisa J. F. Houwink. 2021. [Modernizing family health history: achievable strategies to reduce implementation gaps](#). *Journal of Community Genetics*, 12(3):493–496.
- Kecheng Zhan, Weihua Peng, Ying Xiong, Huhao Fu, Qingcai Chen, Xiaolong Wang, and Buzhou Tang. 2021. [Novel Graph-Based Model With Biaffine Attention for Family History Extraction From Clinical Text: Modeling Study](#). *JMIR Medical Informatics*, 9(4):e23587.
- L. Zhou, Y. Lu, C.J. Vitale, P.L. Mar, F. Chang, N. Dhopeswarkar, and R.A. Rocha. 2014. [Representation of Information about Family Relatives as Structured Data in Electronic Health Records](#). *Applied Clinical Informatics*, 5(2):349–367.

## A Annotation Guidelines

**Objective.** The goal of this protocol is to annotate family history details in clinical conversational narratives to support the development and evaluation of family history extraction tools.

### A.1 Excel annotation sheet format

Annotations are recorded in an Excel sheet.

Column	Header (Excel)	Description
A	Document ID	Document identifier from the note header. Used to link annotations to the source note.
B	FamilyMember OR Observation	Top-level entity for the row: FamilyMember, Observation, or NaN (if no family history mentioned in the note).
C	FamilyMember Type	The relative type when Column B = FamilyMember. May also contain numbering to disambiguate multiple relatives of the same type (e.g., "uncle #2").
D	LivingStatus OR SideoftheFamily OR Observation OR Age	Secondary entity linked to the FamilyMember when Column B = FamilyMember.
E	LivingStatus OR SideoftheFamily	The value for Column D (Alive/Dead or Maternal/Paternal).
F	Description of Previous Entities (i.e., observation description, age of onset...)	Free text field for the condition name and relevant details REQUIRED for all Observation rows.
G	AgeofOnset or Age-ofDeath	The value (ranges allowed) of age of onset or age of death from an Observation for a FamilyMember
H	Negated	True if the Observation is not present
I	CauseofDeath	True if the Observation is the cause of death of the FamilyMember

Table 7: Excel columns used for annotation.

Each row corresponds to at least one annotated *Observation*, with columns used to link (i) a document identifier, (ii) an optional family member mention, and (iii) one or more associated attributes and observations. The annotation procedure is as follows.

## A.2 Annotation workflow

**Step 1: Column A - Document identifier.** In Column A, record the Document ID from the header of each note. This identifier links all annotation rows to the source document.

**Step 2: Column B - Select the top-level entity.** In Column B, select the major entity type to be described in the row: *Observation* or *FamilyMember*. An *Observation* is a family-history-relevant condition mention that does not specify an associated family member, for example:

“Patient has no family history of atrial fibrillation.”

If a specific family member is mentioned, select *FamilyMember*. If there is no family history mentioned anywhere in the note, use the *NaN* dropdown option.

**Step 3: Column C - Populate FamilyMember type.** If Column B is *Observation*, specify the observation in Column F (Description of Previous Entities).

If Column B is *FamilyMember*, specify the family member mentioned in Column C (FamilyMember Type), which is suggested to be one of the following: Great-grandfather, great-grandmother, grandmother, grandfather, father, mother, aunt, uncle, cousin, sister, brother, daughter, son, grandson, granddaughter, sibling, niece, or nephew, child, sibling. Please enter other relationships as needed, if they are present in the text. If multiple family members of the same type are mentioned (e.g., multiple uncles), add numbering in Column E (e.g., “uncle #2”).

**Step 4: Column D - Add associated rows for each FamilyMember-Observation pair OR just Observation.** If Column B is *FamilyMember*, capture three additional entity types as separate rows associated with the same FamilyMember Type: *LivingStatus*, *SideoftheFamily*, and *Observation*. Each associated unique Observation entity for a given FamilyMember-Observation pair should be entered on its own row (e.g., if a sister is alive and has two conditions, create two *Observation* rows for a FamilyMember of Sister and record *LivingStatus* in both rows). If general good health is mentioned, it may be recorded as an *Observation*. If *SideoftheFamily* is not mentioned (e.g., “two of my aunts had cancer” without maternal/paternal), do not add *SideoftheFamily* information.

Notes may include abbreviations for family members. Pronouns are resolved to the previously mentioned family member when applicable. For example:

“Her mother had coronary artery disease and hypertension. She also died from sudden cardiac death.”

Here, *She* refers to *mother*.

**Step 5: Fill Column D/E for LivingStatus or SideoftheFamily.** If Column D is *LivingStatus*, Column E should be *Dead* or *Alive*.

*LivingStatus* may be inferred from context when not explicitly stated; for example:

(1) “cousin has stomach cancer”

→ *LivingStatus* = *Alive*;

(2) “cousin undergoing treatment for stomach cancer”

→ LivingStatus = Alive;

(3) “cousin passed away from stomach cancer”

→ LivingStatus = Dead.

If Column D is *SideoftheFamily*, Column E should be *Paternal* or *Maternal*.

### Step 6: Record Observation details in Column F.

For any *Observation* (whether selected in Column B or Column D), Column F (Description of Previous Entities) is REQUIRED. Use this free text field to specify the condition name and any relevant details. For all entity types, Column F (Description of Previous Entities) is used to record details. Distinguish missing/unknown information from explicitly absent information: *Observation* – Unknown, indicates missing information or uncertainty (e.g., “I don’t know”); *Observation* – None, indicates not relevant/not pertinent/not significant when explicitly stated or asked about.

**Step 7: Columns G, H, and I - Modifiers, Negated, and Cause of Death** Column G (AgeofOnset or AgeofDeath): Record the numeric value (or range) of the age of onset or age of death related to an *Observation*.

Column H (Negated): For negative or absent family history (e.g., “No history of breast cancer”), record the relevant *FamilyMember* (if present; otherwise select *Observation*). Specify the condition in Column F, and mark Column H as True. Distinguish missing/unknown information from explicitly absent information: *Observation* – Unknown indicates missing information or uncertainty (e.g., “I don’t know”); *Observation* – None indicates not relevant/not pertinent/not significant when explicitly stated or asked about.

Column I (CauseofDeath): If an *Observation* is explicitly stated as the cause of the family member’s death, mark Column I as True. If an *Observation* can be inferred as the cause of death based off the phrased order (i.e. mention of death and mention of condition within the sentence or following sentence), mark Column I as True. For example:

“Well, doctor, he passed away. He had mesothelioma.”

→ Column I = True (for the mesothelioma *Observation*).

## A.3 Annotation Example

Excerpt from document 945

**Doctor:** Can you tell me more about your family history, ma’am? How is your father doing?

**Patient:** Well, doctor, my father passed away from leukemia a while back.

Column	Field	Value
A	doc_id	945
B	FamilyMember	Father
C	AgeofOnset	–
D	Observation	leukemia
E	SideoftheFamily	–
F	LivingStatus	Dead
G	Age	–
H	AgeofDeath	–
I	CauseofDeath	TRUE
K	Negated	–

Table 8: Example annotation from document 945.

## B Error analysis between annotators

The following section details error analysis for annotations in both datasets.

Table 9 details the disagreements found after the first pass of annotations (before adjudication). Annotators are denoted by A and B. Semantic disagreement examples can be seen in Table 10.

Category	KIT	MTS
Exact/Boundary Match	629 (87.6%)	181 (67.8%)
Semantic Disagreement	31 (4.3%)	3 (1.1%)
Missing in B	32 (4.5%)	48 (18.0%)
Extra in B	26 (3.6%)	35 (13.1%)

Table 9: Annotator agreement breakdown for *Observations* and *FamilyMembers*. Missing in B means the mention was found extra within A, and extra in B signifies the mention was missing from B. KIT: n=718, MTS: n=267.

The examples in Table 10 were considered errors under a strict match policy but were considered lenient matches. In the final dataset, differences were adjudicated through discussion and removing extraneous information:

## C Model prompt and FHx schema

### C.1 Schema Definition

We define two Pydantic classes for structured family history extraction. The `FHx_Object` class represents a single family history entry with the fol-

#	Ann. A	Ann. B	Final
1	FM: Uncle Obs: had polio, leg deformities	FM: Uncle Obs: polio, leg deformities	FM: Uncle Obs: polio, leg deformities
2	FM: Father LS: Dead Age: 40s	FM: Father LS: Dead Age: died in 40s	FM: Father LS: Dead Age: 40s
3	FM: Brother LS: Alive unknown	FM: Brother LS: Alive	FM: Brother LS: Alive

Table 10: Lenient matches with final annotations. FM = *FamilyMember*, Obs = *Observation*, LS = *LivingStatus*

lowing fields: reasoning (string), FamilyMember (optional string), AgeofOnset (optional integer), Observation (optional string), SideoftheFamily (optional literal: “Maternal”, “Paternal”, or “Unknown”), LivingStatus (optional literal: “Alive”, “Dead”, or “Unknown”), Age (optional integer), AgeofDeath (optional integer), CauseofDeath (optional boolean), CUI (optional string), and Negated (optional boolean). The FamilyHistory class contains a list of FHx\_Object instances called full\_history.

## C.2 Base System Prompt

**Role:** You are an assistant that is reviewing various texts to extract the relevant family health history information for several different condition groups.

**Scope:** Proband’s relatives only. Exclude the proband’s own conditions and any partner/spouse and their relatives.

**Output:** One FamilyHistory object with a list of FHx\_Objects. Each FHx\_Object may represent:

- A condition (Observation) linked to a specific family member, or
- A condition with no specific member (FamilyMember = null), or
- A family member record without a condition.

Fill null for any field not explicitly stated. Include negative and unknown family history.

**FamilyMember relation fields:** LivingStatus (alive/dead/unknown), Age (integer if explicit), AgeOfDeath (integer if explicit), SideOfFamily only for second-/third-degree relatives when explicitly stated.

**Observation relation fields:** CUI (UMLS target concept; null if uncertain), AgeOfOnset (integer if explicit), CauseOfDeath = true only if the text explicitly says the condition caused death; otherwise false.

**Negated field:** boolean; set to true ONLY if the text explicitly states a negative family history or that the member is healthy.

Provide concise rationale for extracted FHx\_Object and reasons behind CUI mapping in the reasoning field.

### Rules:

1. Extract only explicitly stated facts; otherwise use null. Do not infer ages, sex, side, or subtypes.
2. **Negateds:** If a text explicitly states a negative family history (e.g., “no history of cancer” or “brother is healthy”), extract the base condition in the Observation field (e.g., “cancer”) and set the Negated field to true. For general healthy statements, set Observation to “any conditions” and Negated to true. Do not write the word “no” in the Observation field.
3. De-duplicate by member + condition (CUI) + status. Have a unique FHx\_Object per condition per member.

## C.3 Differences between Zero-shot and Few-shot approaches

The few-shot in-context learning approach provided one example from both *FHexchange* sub-datasets. More details regarding the examples can be found in the dataset repository on Github.

## D Baseline Experiment Results and Error Analysis

Model	Setting	FM	Obs.	FM-Obs.	CUI	Obs.-CUI
GPT-4o	ZS	.994 (.99/1.0)	.578 (.53/.64)	.632 (.66/.61)	.354 (.36/.35)	.229 (.23/.23)
	FS	.986 (.98/.99)	.638 (.58/.71)	.702 (.71/.69)	.373 (.37/.38)	.258 (.23/.29)
.000						
GPT-5-mini	ZS	.972 (.97/.97)	.726 (.66/.81)	.824 (.85/.80)	.115 (.48/.07)	.100 (.42/.06)
	FS	.891 (.97/.82)	.649 (.63/.67)	.711 (.79/.65)	.255 (.44/.18)	.189 (.37/.13)
Llama 3.3	ZS	.077 (1.0/.04)	.000	.000	.000	.000
	FS	.201 (.93/.11)	.101 (.83/.05)	.100 (.86/.05)	.107 (.11/.11)	.010 (.08/.01)

Table 11: Complete performance on *FHexchange-KIT*. Values: F1 (Precision/Recall). FM = *FamilyMember*, Obs. = *Observation*, ZS = Zero-Shot, FS = Few-Shot.

Model	Set	FM	Obs.	FM-Obs.	CUI	Obs.-CUI	Neg.	Obs.-Neg.	Living	FM-LS	Side	FM-Side
GPT-4o	ZS	.851 (.97/.76)	.546 (.64/.48)	.528 (.61/.47)	.597 (.61/.59)	.307 (.36/.27)	.983 (.98/.99)	.490 (.55/.44)	.681 (.66/.70)	.617 (.66/.58)	.400 (.67/.29)	.500 (1.0/.33)
	FS	.873 (.97/.79)	.692 (.79/.62)	.695 (.74/.66)	.564 (.56/.57)	.386 (.42/.36)	.982 (.97/.99)	.269 (.50/.18)	.641 (.59/.70)	.607 (.61/.60)	.133 (1.0/.07)	.105 (1.0/.06)
GPT-5-mini	ZS	.774 (.79/.76)	.674 (.75/.61)	.625 (.67/.59)	.234 (.80/.14)	.106 (.47/.06)	.984 (.98/.99)	.612 (.65/.58)	.740 (.80/.69)	.613 (.70/.55)	.429 (.60/.33)	.345 (.46/.28)
	FS	.881 (.94/.83)	.670 (.74/.61)	.658 (.68/.64)	.553 (.63/.49)	.371 (.49/.30)	.980 (.97/.99)	.235 (.46/.16)	.705 (.69/.72)	.680 (.71/.65)	.000	.000
Llama 3.3	ZS	.043 (.83/.02)	.000	.000	.000	.000	1.00 (1.0/1.0)	.000	.889 (1.0/.80)	.038 (1.0/.02)	.000	.000
	FS	.568 (.93/.41)	.468 (.78/.33)	.478 (.74/.35)	.405 (.39/.42)	.188 (.30/.14)	.992 (.98/1.0)	.169 (.50/.10)	.688 (.64/.75)	.452 (.64/.35)	.222 (.50/.14)	.174 (.40/.11)

Table 12: Complete performance on *FHexchange-MTS*. Values: F1 (Precision/Recall). FM = *FamilyMember*, Obs = *Observation*, Neg. = *Negated*, LS = *LivingStatus*. ZS = Zero-Shot, FS = Few-Shot. Age-related attributes were 0.000 across all models and omitted.

Observation	GPT-4o Reasoning	GPT-5-mini Reasoning
Testicular cancer (Father)	“User confirmed the father has been diagnosed with testicular cancer.” CUI field: C0039590	“Father was indicated to have had a cancer described by the user as ‘testicular’... CUI set to null because a precise UMLS mapping was not provided/confirmed.” CUI field: null
Cataracts (Mother)	“Based on the selection indicating the mother has had hearing and/or eye conditions, further specifying cataracts.” CUI field: C0007169*	“The user explicitly selected ‘Mother’ as having a hearing/eye condition... CUI is set to null because a specific UMLS concept mapping was not provided and is uncertain.” CUI field: null
Heart attack (Father)	“User confirmed the father has experienced a heart attack.” CUI field: C0027051	“Father was selected as having a heart and blood condition... CUI set to null because a precise UMLS mapping was not provided/confirmed.” CUI field: null

Table 13: Analysis of model behavior through schema *reasoning* field with CUI. GPT-4o assigns UMLS codes to *Observations* at a higher rate than GPT-5-mini (81.3% versus 10.4% assignment rate, respectively), but contains erroneous mappings and hallucinations\*. For example, C0007169 is not a valid CUI. Few-shot prompting increases GPT-5-mini’s CUI F1 from 0.115 to 0.255 (*FHexchange-KIT*) and 0.234 to 0.553 (*FHexchange-MTS*).

Doc	Predicted	Gold
<i>FHex-MTS</i>		
735	aunt, type 2 diabetes mellitus	aunt, type 2 diabetes
1161	grandmother, tachycardia	grandmother, fast heart beat
945	mother, kidney failure	mother, renal failure
<i>FHex-KIT</i>		
75	grandparent, irritable bowel syndrome	grandparent, irritable bowel syndrome (ibs)
86	father, leukemia	father, blood cancer
86	grandparent, renal cancer	grandparent, kidney cancer
57	sibling, substance abuse	sibling, drug use disorder

Table 14: *Observation* extraction errors showing semantic mismatches between model predictions and gold annotations.