Overview of the BioCreative/OHNLP 2018 Family History Extraction Task

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

***Abstract*— As a risk factor for many diseases, family history captures both shared genetic variations and living environments among family members. Though there are several systems focusing on family history extraction (FHE) using natural language processing (NLP) techniques, the evaluation protocol of such systems has not been standardized. The BioCreative/OHNLP 2018 Family History Extraction Task aims to encourage the community efforts on a standard evaluation on FHE from clinical narratives. This novel shared task composes two subtasks. Subtask 1 focuses on identifying family member entities and clinical observations (diseases), and Subtask 2 expects the association the living status, side of the family and clinical observations to family members to be extracted. Subtask 2 is an end-to-end task which is based on the result of Subtask 1. We manually curated the first synthetic clinical narrative from family history sections of clinical notes at Mayo Clinic Rochester, the content of which are highly relevant to patients' family history. Five teams have submitted their system predictions in the Subtask 1 and three teams participated the Subtask 2. The best performed run achieved F1-scores of 0.8861 and 0.5708 in Subtask 1 and 2, respectively.**

Keywords—family history extraction; information extraction; natural language processing; named entity recognition; relation extraction.

I. Introduction

As a risk factor for many diseases, family history (FH) captures shared genetic variations among family members [1]. Information such as age, gender, and degree of relatives are also considered when taking into the account of risk assignment of a large number of common diseases. For example, the risk assessment of hypertrophic cardiomyopathy considers one or more first-degree relatives with history of sudden cardiac death under age 40 as a significant factor of sudden cardiac death risk in hypertrophic cardiomyopathy patients [2].

The fact that many care process models use FH information highlights the importance of FH in the decision-making process of diagnosis and treatment. However, acquiring accurate and complete FH information remains challenging for clinical natural language processing (NLP) community. One of the major sources of FH data is from Patient Provide Information (PPI) questionnaires, which is usually stored in semi-structured/unstructured format in electronic health records [3]. In order to provide a comprehensive patient-provided FH data to physicians, there is a need for NLP systems that are able to extract FH from text. Elements of FH data are not pre-determined or limited. They depend on pieces of information that provided by patients about their relatives' health situation during visits. The FH elements may include: disease, family member, cause, medication, age of onset, length of disease, etc. This variety of FH elements makes the extraction process from unstructured data challenging.

The application of NLP methods and resources to clinical and biomedical text has received growing attention over the past years [4], but progress has been limited by difficulties to access shared tools and resources, partially caused by patient privacy and data confidentiality constraints. Efforts to increase sharing and interoperability of existing resources are needed to facilitate the progress observed in the general NLP domain. Leveraging our research in corpus analysis and de-identification research, we have created multiple synthetic data sets for a couple of NLP tasks based on real clinical sentences. In this document, we describe the dataset generated for family history extraction (FHE) from unstructured data.

Though there are several systems focusing on FHE using NLP techniques [5]–[7], there are little efforts on standardize the evaluation protocol of such systems, which makes it challenging for users to compare the performance of different FHE systems. To address this issue, we organized this novel shared task to encourage the community to propose and develop FHE systems.

II. The Sythetic Family History Extraction Corpus

The patient notes we used to curate the corpus are randomly sampled from Mayo Employee and Community Health cohort. We extracted the section titled "family history" in this corpus as the first stage of text selection. Then, we have excluded automatically generated semi-structured texts and sections with extensive social history. As a result, the clinical texts in the corpus focus on patient FH information.

We annotated the corpus using Anafora, a web-based annotation tool for texts [8]. In total of 11 people are involved in the annotation process. 10 of them did the first round of annotation, and each document was annotated by 2 people. One senior study coordinator worked as adjudicator to resolve disagreements from the two annotations.

An example of the entity annotation is shown in Figure 1. The sentence "the patient's maternal grandmother was diagnosed with multiple sclerosis at age 59 and passed away at age 80" are annotated with entities of family members, observation, living status and ages. The incremental ID field of entities is used to distinguish multiple individuals. In this example, we only have one individual under the family member of "maternal grandmother", so all the IDs are 1. The annotation schema of the FHE corpus is illustrated in Figure 2. The corpus is annotated with the following entities and attributes.

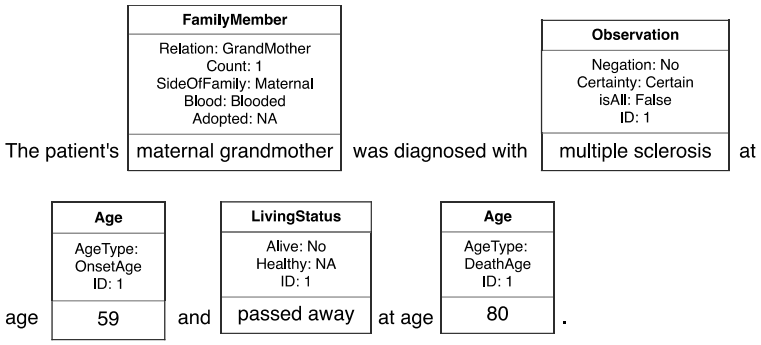

Figure 1 Example entity annotation in FHE corpus

1) Family Members

In this study, we annotated only first and second relatives by blood. The spouses were not considered as blood relatives, thus were excluded from the annotation.

Each family member has several properties:

- Side of Family: (maternal or paternal). The family side mentions are also included in the family member entity annotations.
- Count: the total number of family members under the family member category.
- Blood: whether the family members are fully blood related. For instance, a step-sister with shared mother of the patient is considered as "half-blooded". The default value is "NA" for most of the family members.
- Adopted: whether the family members are adopted to the family.

2) Observation

Any health-related problem including diseases, smoking, suicide, and drinking, excluding auto accident, surgery and medications. The observation entities have several attributes: negation, certainty, whether the observation applies to all family members, and an integer identifier of family member in case, there are more than one person in that family category. The negated observations will have a negation field value of "Yes".

3) Age

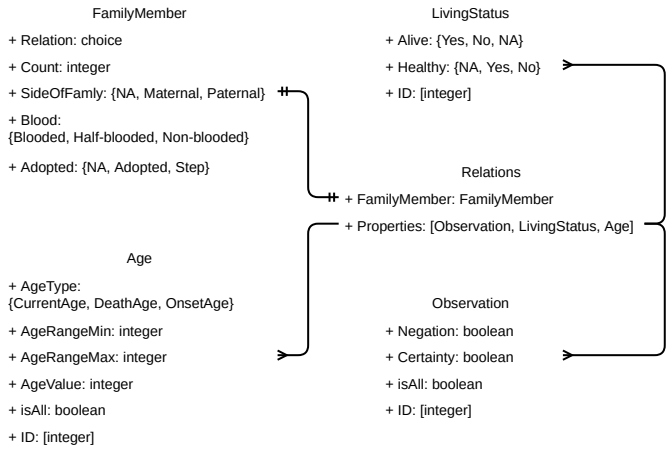

Figure 2 FHE annotation schema

The age mentions related to family member, observation, or death are annotated. The word "age" is not annotated in the age mentions. For ranges of age such as "80s", range min and max values are also annotated.

4) Living Status

Living status are the words and phrases which show health status of the family members. The default value is "Alive: yes" and "Healthy: NA".

All the entities related to a family member category are linked into one chain. In the example shown in Figure 1, the chain has family member of maternal grandmother, and the rest of the chain links other entities related to the family member category. If the patient has multiple family members in the same category (e.g. several brothers), all the entities related to any of the brothers will be linked into a chain of "Brother". The entities can be later restored to each individual family member by their IDs. The incremental IDs are annotated to identify observation, age, and living status from different individuals within the same category.

As part of the annotation process, we de-identified the dataset. All the patient protected information, such as names, locations and identifiers, has been deidentified according to the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule [9]. In addition, the observations, family members and ethnicities are shuffled among the whole corpus. The numeric fields like dates and phone numbers and IDs are manually replaced with synthetic strings. As a result, the corpus should only be used for studies of information extraction purposes, rather than clinical applications or statistics requiring clinical relevance.

We have included in total of 99 documents in the training set and 50 documents for test. The training set was released to participants containing both text and annotation files, while for the test set, only the raw text files were released. Some statistics on the corpus is shown in Table I.

III. Evaluation

For the entity identification subtask (Subtask 1), the participants are expected to provide two types of information: family members mentioned in the text and the observations (diseases) in the family history. We only used normalized family members for evaluation. The normalized family members are listed in Table II.

Table I Corpus Statistics

	Train	Test
Document	99	50
Family Member	802	331
Age	756	289
Living Status	415	181
Observations	978	465
Chains	651	280
Average chain length	2.92	2.84

Only the occurrence in the document is evaluated. For family member entities appearing multiple times in a document, only one true positive is counted. For first degree relatives (e.g. parents, children, siblings), the side of family should always be "NA". To reduce the ambiguity in observation extraction, we accept partial matching of the observations. For example, an extraction of "diabetes" in the phrase of "type 2 diabetes" will be considered as a true positive when calculating F1 score. We limited the submissions of observations to no more than 4 tokens to avoid the abuse of the flexibility. To reduce the complexity of the task, the negation information is removed from evaluation for both subtasks.

Table II Normalized Family Members

Degree of family members	Normalized family members
1	Father, Mother, Parent, Sister, Brother, Daughter, Son, Child
2	Grandmother, Grandfather, Grandparent, Cousin, Sibling, Aunt, Uncle

In the Subtask 2, the participants need to provide summarized information between family members and observations. For family members, the participants are asked to provide a tuple of (Family member, side of family, living status coding). For the observation extraction, the systems are asked to provide tuples of (Family member, side of family, observation). In cases where there are more than one observation for one family member category, separate tuples are expected.

We use only one score to represent living status for each individual. For the simplicity of comparison, we encoded the two fields of living status (alive, healthy) into one integer. For both "Alive" and "Healthy" properties, the results of "Yes", "NA", "No" are encoded as 2, 1, 0, respectively. The living status score is the alive score multiplying the healthy score. For example, for a family member with "Alive" as "Yes" and "Healthy" as "Yes", the living status score should be 2 * 2 = 4. For a family member with "Alive" as "No" and "Healthy" as "NA", the living status score should be 0 * 1 = 0. Therefore, the higher of the encoded living status value, the better the family member's current condition is. There are also cases that multiple relatives under the family member category, (e.g. multiple maternal aunts) having different living status scores. In such cases, the minimum of the scores as the final score for that category.

To be considered as a correct prediction (true positive) for family members, all of the fields have to be matched, including living status. For Subtask 2, the observation matching criterion is the same as Subtask 1, where partial matching is allowed. Observations applied to all relatives should not be included. For example, in the sentence "there were no reports of mental illness", the observation of "mental illness" should not appear in any family members.

We use standard F1-score as the evaluation (ranking) metrics. Specifically,

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2\ Precision \cdot Recall}{Precision + Recall}$$

where true positive (TP) denotes the number of correct predictions, false positive (FP) denotes the number of system predictions that do not exist in the gold standard, and false negative (FN) denotes the number of gold standard records that do not exist in the system predictions. More details on the evaluation and the evaluation script can be found at https://github.com/ohnlp/fh_eval. The inter annotator agreement (IAA) between two annotators measured before the de-identification process in F1 scores are 0.8379 and 0.7012 for Subtask 1 and 2, respectively.

IV. Results

A total of 5 teams submitted overall 14 submissions for the FHE task. The detailed descriptions of the participated systems can be found in the BioCreative proceedings. Table III shows the number of submissions by each team.

Table III Participating teams and submissions

Team	Subtask 1: entity identification	Subtask 2: family history extraction
NTTU [10]	3	
BIOB [11]	1	1
NLPHard [12]	1	1
JE_NLP [13]	1	
HIT [14]	3	3
Total	9	5

Each team can submit up to 3 runs for each subtask (i.e. Subtask 1 and 2). The final ranking of Subtask 1 and 2 are shown in Table VI and Table V, respectively. For the entity identification subtask, we received 9 runs from 5 teams. The best run overall is Run 2 from team HIT, which achieved the F1 score of 0.8861. It is interesting that 5 out of the 9 runs have achieved higher F1 scores than the IAA of 0.8379. This may be

Table VI Results for Subtask 1: Entity Identification

Team	Run	Family Member			Observation			Overall		
		P	R	F1	P	R	F1	P	R	F1
NTTU	1	0.8779	0.8303	0.8534	0.8024	0.8944	0.8459	0.8285	0.8698	0.8486
	2	0.8375	0.8375	0.8375	0.8073	0.8944	0.8486	0.8182	0.8726	0.8445
	3	0.7914	0.7942	0.7928	0.8251	0.9011	0.8614	0.8128	0.8601	0.8358
BIOB	1	0.7232	0.8773	0.7928	0.8481	0.8157	0.8316	0.7932	0.8393	0.8156
NLPHard	1	0.8659	0.8159	0.8401	0.8926	0.7843	0.8349	0.8819	0.7964	0.837
JE_NLP	1	0.9259	0.361	0.5195	0.6233	0.6247	0.624	0.6823	0.5235	0.5925
HIT	1	0.9008	0.852	0.8757	0.8589	0.9169	0.887	0.8738	0.892	0.8828
	2	0.8806	0.852	0.8661	0.8933	0.9034	0.8983	0.8886	0.8837	**0.8861**
	3	0.881	0.8556	0.8681	0.8857	0.9056	0.8956	0.884	0.8864	0.8852

Table V Results for Subtask 2: Family History Extraction

Team	Run	Family Member			Observation			Overall		
		P	R	F1	P	R	F1	P	R	F1
BIOB	1	0.4785	0.6211	0.5405	0.5684	0.5	0.532	0.5304	0.5402	0.5352
NLPHard	1	0.5652	0.5652	0.5652	0.6913	0.4907	0.574	0.6394	0.5155	**0.5708**
HIT	1	0.6378	0.5031	0.5625	0.6015	0.5031	0.5479	0.6131	0.5031	0.5527
	2	0.5462	0.4037	0.4643	0.511	0.358	0.4211	0.5231	0.3732	0.4356
	3	0.5726	0.441	0.4982	0.5084	0.3735	0.4306	0.5304	0.3959	0.4534

due to the improvement of the annotation quality after the adjudication process, so that a properly tuned systems can outperform human in identifying family history related concepts.

For Subtask 2, family history extraction subtask, we received 6 runs from 3 teams. The top team, NLPHard, has achieved the F1 score of 0.5708 overall. From the best run performance, we can conclude that the Subtask 2 is more difficult comparing to the Subtask 1. This is also supported by the performance gap between the IAAs of Subtask 1 and 2. The reasons of this are two-fold. First, as an end-to-end relation extraction task, this is partially because the relation predictions need to be based on the prediction results of Subtask 1 on family members, side of family and observations. Errors in the entity extraction tasks will pass to the relation extraction task, causing errors in predicting the observations and family member living status. Second, from previous studies on end-to-end relation extraction tasks, the performance in relation extraction tasks are lower than named entity recognition tasks [15], [16]. A successful system also needs to consider co-reference resolution, which could be considered as a standalone task for NLP systems [17].

V. Conclusions and Future work

The FHE corpus and shared task has successfully encouraged participants internationally to contribute to FHE systems for clinical narratives. The synthetic corpus curation process and the FHE system evaluation for this shared task can leverage open science in the biomedical especially clinical NLP community. The corpus can be valuable in encouraging the development of clinical NLP systems on FHE, and further fostering secondary usage of unstructured EHR data in precision medicine.

In future, we will organize the shared task in different perspectives, including providing gold standard entity annotations to encourage participants focusing on the entity association component. We will also include more challenging tasks, such as evaluating systems by person rather than family member categories, and including more entities and attributes in the evaluation. We would also like to increase the number of documents by annotating more documents, preferably from other institutions, so that more comprehensive patterns can be trained and evaluated. The annotation schema of this shared task will also be integrated with other common data models.

ACKNOWLEDGMENT

We would like to thank the FHE dataset annotators: Donna Ihrke, Xin Zhou, Suyuan Peng, Jun Jiang, Nan Zhang. This task is made possible by National Institute of Health, National Institute of General Medical Sciences R01-GM102282 and National Center for Advancing Translational Sciences U01TR02062. The student travel grants and awards are sponsored by NIH grant 5R01GM080646-12.

REFERENCES

[1] J. J. McCarthy and B. A. Mendelsohn, "Family History," in *Precision Medicine: A Guide to Genomics in Clinical Practice*, New York, NY: McGraw-Hill Education, 2016.

[2] ESC, "2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy," *Eur. Heart J.*, 2014.

[3] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Liu, F. Shen, and H. Liu, "Systematic Analysis of Free-Text Family History in Electronic Health Record," *AMIA Summits Transl. Sci. Proc.*, vol. 2017, pp. 104–113, 2017.

[4] Y. Wang *et al.*, "Clinical information extraction applications: A literature review," *J. Biomed. Inform.*, vol. 77, 2018.

[5] R. Bill, S. Pakhomov, E. S. Chen, T. J. Winden, E. W. Carter, and G. B. Melton, "Automated extraction of family history information from clinical notes.," *AMIA Annu. Symp. Proc.*, vol. 2014, pp. 1709–17, 2014.

[6] N. Lewis, D. Gruhl, and H. Yang, "Dependency parsing for extracting family history," in *Proceedings - 2011 1st IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology, HISB 2011*, 2011, pp. 237–242.

[7] S. Goryachev, H. Kim, and Q. Zeng-Treitler, "Identification and extraction of family history information from clinical reports.," *AMIA Annu. Symp. Proc.*, pp. 247–51, 2008.

[8] W.-T. Chen and W. Styler, "Anafora: A Web-based General Purpose Annotation Tool," *Proc. 2013 NAACL HLT Demonstr. Sess.*, no. June, pp. 14–19, 2013.

[9] "The HIPAA Privacy Rule." [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy.

[10] F. Wang, C. Wang, and H. Dai, "Family History Information Extraction with Neural Seqeunce Labeling Model," in *BioCreative 2018 Workshop Proceedings*, 2018.

[11] D. Kim, S. Shin, H. Lim, and S. Kim, "Efficient memory-based approaches for tagging named entities and relations in clinical text," in *BioCreative 2018 Workshop Proceedings*, 2018.

[12] . A., V. Gela, and S. Madgi, "Hybrid Approach for End-to-End Entity Recognition and Entity Linking using CRFs and Dependency Parsing," in *BioCreative 2018 Workshop Proceedings*, 2018.

[13] E. Tseng and J. Lee, "A combined rule-based and statistical approach to family history extraction from unstructured text," in *BioCreative 2018 Workshop Proceedings*, 2018.

[14] X. Shi *et al.*, "Family History Information Extraction Via Joint Deep Learning," in *BioCreative 2018 Workshop Proceedings*, 2018.

[15] I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo, "Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)," *Proc. Seventh Int. Work. Semant. Eval. (SemEval 2013) vol. 2, Assoc. Comput. Linguist.*, vol. 2, no. SemEval, pp. 341–350, 2013.

[16] C. H. Wei *et al.*, "Assessing the state of the art in biomedical relation extraction: Overview of the BioCreative V chemical-disease relation (CDR) task," *Database*, vol. 2016, 2016.

[17] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, and B. R. South, "Evaluating the state of the art in coreference resolution for electronic medical records.," *J. Am. Med. Inform. Assoc.*, vol. 19, no. 5, pp. 786–91, 2012.