User Manual

September, 2021

1 Introduction

E-Pedigrees: a large-scale automatic family pedigree prediction application which was developed as a novel fully automated software to construct family pedigrees from information readily available in an EHR system [Huang et al., 2021]. *E-pedigrees* infers familial relationships using two previously published prediction algorithms including Family Pedigree Prediction Algorithm **(FPPA)** [Huang et al., 2017] and Relationship Inference from the Electronic Health Record **(RIFTEHR)** [Polubriaginof et al., 2018].

2 Requirements

python 3.8+.

networkx 2.6+. If you use older version of networkx, the functions of "connected_component_subgraphs(G, copy=True)" and "nodes()" may have been deprecated in your version.

3 Usage:

Please follow the exact input format for all the input files. You can choose to run "FPPA", "RIFTEHR", or "BOTH", a combination of FPPA and RIFTEHR. Besides, you can provide your own reliable pedigrees file in the PED format for constructing new family pedigrees with existing population.

3.1 FAAP:

command-line:

run *E-pedigrees*: python main.py FPPA

Enter your input files for FPPA: address.csv name.csv demo.csv account.cs pc.txt familyTree.csv

Enter one PED file if any: ped.csv [optional] (leave it blank if you do not have a PED file)

3.2 RIFTEHR

command-line:

run *E-pedigrees*: python main.py RIFTEHR

Enter your input files for RIFTEHR: patient.csv ec.csv familyTree.csv

Enter one PED file if any: ped.csv [optional] (leave it blank if you do not have a PED file)

3.3 Both:

command-line:

run *E-pedigrees*: python main.py both

Enter your input files for BOTH: address.csv name.csv demo.csv account.cs pc.txt patient.csv ec.csv familyTree.csv

Enter one PED file if any: ped.csv [optional] (leave it blank if you do not have a PED file)

4 Input Files

4.1 FPPA

Input files for address file in table 1, name file in table 2, demographic file in table 3, account file in table 4.

study_id	street_1	street_2	city	state	zip	from_year	thru_year
1	790393		7200	28	18216		
10	117141		5115	28	11753		2005
56	221591	448275	2893	28	9427	2003	2011

Table 1: Address information file format.

study_id	last_name_id	first_name_id	middle_name_id	from_year	thru_year
1	103775	53806			
10	46972	44623		2005	2011
50	2696	62099		1997	2007
50	105616	62099			1997

Table 2: Name information file format.

study_id	gender_code	birth_year	deceased_year	PHONE_NUM_id	from_year	thru_year
1	F	1989				
2	F	1947		134271		2011
282056	U	1986	2010			

Table 3: Demographic information file format.

study_id	ACCT_NUM_id	from_year	thru_year
2	982162		2011
10	523063	2005	2011

Table 4: Account information file format.

4.2 RIFTEHR

Input files for patient file in table 5, and emergency contact file in table 6.

PatientID	FirstName	LastName	Sex	PhoneNumber	Zipcode	birth_year	deceased_year
1	103775	53806	M	1112223333	18216	1970	
10	46972	44623	M	2223334444	11753	1972	
50	2696	62099	F	3334445555	18216	1980	
96	105616	53806	F	1112223333	10032	1956	
122	345228	44623	F	2223334444	11753	1990	

Table 5: Patient information file format.

PatientID	EC_FirstName	EC_LastName	EC_PhoneNumber	EC_Zipcode	EC_Relationship
1	105616	53806	1112223333	18216	Mother
10	345228	44623	2223334444	11753	Father

Table 6: Emergency contact file format.

4.3 PED file

Pedigree file format in table 7.

family_ID	num_fam_member	individual_ID	Maternal_ID	Paternal_ID	Gender
1	5	50	1112223333	18216	M
2	3	96	2223334444	11753	F

Table 7: Pedigree file format.

5 BOTH

Input files for "BOTH" contains address information file in table 1, name information file in table 2, demographic information file in table 3, account information file in table 4, patient information file in table 5, and emergency contact file in table 6.

6 Output file

Finally, a PED format output file in table 8 will be generated. It contains the family ID, number of family member, individual ID, maternal ID, Paternal ID, and Sex. This output family pedigrees can be used as a cohort for downstream analyses related to family history.

family_ID	num_fam_member	individual_ID	Maternal_ID	Paternal_ID	Gender

Table 8: Family Pedigrees file header format.

References

Xiayuan Huang, Nicholas Tatonetti, Katie LaRow, Brooke Delgoffee, John Mayer, David Page, and Scott J Hebbring. E-Pedigrees: a large-scale automatic family pedigree prediction application. *Bioinformatics*, 37(21):3966–3968, 06 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab419. URL https://doi.org/10.1093/bioinformatics/btab419.

Xiayuan Huang, Robert C Elston, Guilherme J Rosa, John Mayer, Zhan Ye, Terrie Kitchner, Murray H Brilliant, David Page, and Scott J Hebbring. Applying family analyses to electronic health records to facilitate genetic research. *Bioinformatics*, 34(4):635–642, 09 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx569. URL https://doi.org/10.1093/bioinformatics/btx569.

Fernanda C.G. Polubriaginof, Rami Vanguri, Kayla Quinnies, Gillian M. Belbin, Alexandre Yahi, Hojjat Salmasian, Tal Lorberbaum, Victor Nwankwo, Li Li, Mark M. Shervey, Patricia Glowe, Iuliana Ionita-Laza, Mary Simmerling, George Hripcsak, Suzanne Bakken, David Goldstein, Krzysztof Kiryluk, Eimear E. Kenny, Joel Dudley, David K. Vawdrey, and Nicholas P. Tatonetti. Disease heritability inferred from familial relationships reported in medical records. *Cell*, 173(7):1692–1704.e11, 2018. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2018.04.032. URL https://www.sciencedirect.com/science/article/pii/S0092867418305257.