Bioinformatics, 34(4), 2018, 635–642
doi: 10.1093/bioinformatics/btx569
Advance Access Publication Date: 14 September 2017
Original Paper

OXFORD

Genetics and population analysis

Applying family analyses to electronic health records to facilitate genetic research

Xiayuan Huang[1], Robert C. Elston[2], Guilherme J. Rosa[3], John Mayer[4], Zhan Ye[4], Terrie Kitchner[5], Murray H. Brilliant[5,6], David Page[1,7] and Scott J. Hebbring[5,6,*]

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53792, USA, [2]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA, [3]Department of Animal Science, University of Wisconsin-Madison, Madison, WI 53706, USA, [4]Biomedical Informatics Research Center, [5]Center for Human Genetics, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA, [6]Department of Medical Genetics and [7]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on February 24, 2017; revised on June 22, 2017; editorial decision on September 3, 2017; accepted on September 13, 2017

Abstract

Motivation: Pedigree analysis is a longstanding and powerful approach to gain insight into the underlying genetic factors in human health, but identifying, recruiting and genotyping families can be difficult, time consuming and costly. Development of high throughput methods to identify families and foster downstream analyses are necessary.

Results: This paper describes simple methods that allowed us to identify 173 368 family pedigrees with high probability using basic demographic data available in most electronic health records (EHRs). We further developed and validate a novel statistical method that uses EHR data to identify families more likely to have a major genetic component to their diseases risk. Lastly, we showed that incorporating EHR-linked family data into genetic association testing may provide added power for genetic mapping without additional recruitment or genotyping. The totality of these results suggests that EHR-linked families can enable classical genetic analyses in a high-throughput manner.

Availability and implementation: Pseudocode is provided as supplementary information

Contact: HEBBRING.SCOTT@marshfieldresearch.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

An electronic health record (EHR) is a digital representation of a patient's current and past health history that often includes diagnoses (e.g. ICD coding), medications and laboratory test results. With extensive phenotypic data accessible through an EHR, the use of DNA biobanks linked to an EHR has expedited population-based genomics research, such as Genome-Wide Association Studies (GWASs) of unrelated individuals. Although GWASs with or without EHR data have yielded substantial advances in human genomics (Welter *et al.*, 2014), they have inherent limitations. For example, most variants identified by GWAS are intergenic and have associations with weak effect sizes (McCarthy *et al.*, 2008; Manolio *et al.*, 2009; Visscher *et al.*, 2012). The substantial challenges faced by population-based study designs, in combination with advances in whole genome/exome sequencing, have partially led to a revitalization of family-based studies (Chong *et al.*, 2015). Unfortunately, identifying informative families for research can result in ascertainment biases (Emilsson *et al.*, 2015) and is extremely difficult, time consuming and costly.

To address the current challenges in family-based research, the primary aim of this paper is to develop a fully automated approach to construct family pedigrees from information readily available in a typical EHR. The secondary aim is to evaluate the utility of EHR-linked families for genetic research. Hence, analyses of EHR-linked pedigrees have the potential to be more akin to GWAS in the sense of being high-throughput and automated.

To address the aims of creating and evaluating the utility of an automated family prediction method, we designed a hand-constructed decision tree algorithm to predict family pedigrees from Marshfield Clinic's EHR system. We further developed a novel logistic regression for familial relatedness (LRFR) model that measures the correlative relationship between genetic relatedness (Wright, 1922) and phenotypic concordance (Mayer *et al.*, 2014) in a clinical population. Lastly, we demonstrate that families linked to an EHR may have great value in genetic mapping. Given the increasing standardization of EHRs, the methods presented can be easily applied to other EHRs to expedite and provide additional statistical power for genomic research.

2 Materials and methods

2.1 Family pedigree prediction

Marshfield Clinic has a fully integrated EHR system that began in 1984 with ICD9 diagnostic data available since 1979. The Marshfield Clinic EHR provides a pool of medical records for nearly 2.6 million patients. Data of relevance include basic demographic information, such as last name, date of birth, home address, billing account and gender. It is these demographic data, which are often patient-reported, necessary for billing purposes, and likely available in most EHR systems, that were used to predict familial relationships. For protection of patient privacy, all identifiers, such as names, addresses, account numbers and even diagnosis codes, were mapped to positive integers prior to any processing, with the reverse mapping held in a different secure location. Dates were similarly mapped, but with care that time differences, in days, could be computed between any two de-identified dates.

The first elements of the decision tree logic used to construct families were shared home address and shared last name, to identify potential pairs of parent–child or sibling relationships. Because last name and home address may change over time (e.g. due to marriage or children leaving the parental household), we considered only those relationships that shared a common address for over three years. To distinguish the parent–child relationship from siblings, we considered the age difference between each pair (Fig. 1 and Supplementary Fig. S1A). After filtering data and running the decision tree algorithm, we predicted over 500 000 parent–child relationships and over 100 000 sibling relationships from 2.6 million individuals recorded in Marshfield Clinic's EHR.

To further refine the familial relationships, we used parent-child pairs as input, filtering out incorrect multiple parents using match versus mismatch of phone number and billing account attributes. Although we recognize family dynamics and structures can be complex, patients were included only if they had two parents of opposite sex (Fig. 1 and Supplementary Fig. S1B) consistent with our focus on the study of genetic relationships. In the end, only the parent–child relationships were used to construct the pedigrees, though in future work sibling predictions could be used to further refine or extend some pedigrees. After filtering and refining the parent–child relationships, we composed these relationships into family pedigrees using the graphical algorithm package NetworkX (Hagberg et al., 2008). In total,

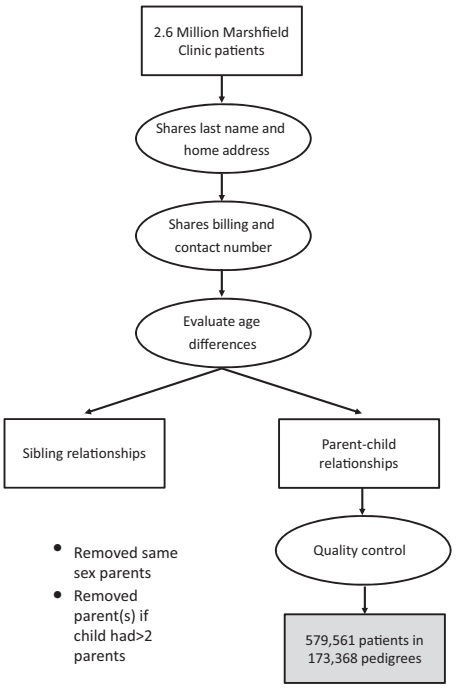


Fig. 1. Flowchart of parent–child relationship decision tree algorithm

173 368 family units with two or more generations were predicted. For each family, an individual from the youngest generation that maximized the number of parents identified was selected as the proband.

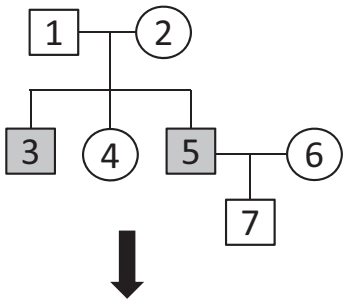
A trained study coordinator conducted manual chart review on 40 randomly selected families (259 total patients) to validate the family prediction algorithm. These families included 20 randomly selected 'standard' and 20 randomly selected 'half-sib' families containing half-siblings—by definition standard families contained no half-siblings. They were assessed by determining the number and percentage of relationships as being true or false positives, and true or false negatives. It was assumed that families with predicted half-siblings may be difficult to identify, prone to errors and thus represent a unique subset to evaluate the accuracy of the prediction algorithm.

2.2 Disease specific studies

2.2.1 LRFR

Phenotypes were extracted from Marshfield Clinic's EHR. Cases with color blindness and muscular dystrophy (MD) were defined by International Classification of Disease, version 9 [ICD9] codes (ICD9 368.5 and 359.1, respectively). An additional 28 broadly defined phenotypes were also evaluated using a 'roll-up' strategy to define affected and unaffected individuals. Specifically, individuals coded for disease specific ICD9 codes were rolled-up into more general ICD9 codes [e.g. 750.27 ('diverticulum of pharynx')→750.2 ('other specified congenital anomalies of mouth and pharynx')→750 ('other congenital anomalies of upper alimentary tract')]. These 28 broadly defined diseases included 19 presumed 'heritable' conditions (congenital codes: ICD9 741–759) and 9 control phenotypes (accidental fall codes: ICD9 E880-E888) that may not have strong heritable influences. ICD9 codes are often entered directly by physicians or administrative staff.

To evaluate the genetic contributions to a particular disease phenotype, we attempted to measure the extent to which inferred



Relationship Pair	Relatedness	Label
3 and 1	0.5	0
3 and 2	0.5	0
3 and 4	0.5	0
3 and 5	0.5	1
3 and 6	0	0
3 and 7	0.25	0

Relationship Pair	Relatedness	Label
5 and 1	0.5	0
5 and 2	0.5	0
5 and 4	0.5	0
5 and 6	0	0
5 and 7	0.5	0

Fig. 2. LRFR. Flowchart of Logistic Regression for Familial Relatedness (LRFR). This figure shows an example family pedigree with two affected individuals (grey). At the bottom, a contingency matrix shows input data for logistic regression

genetic relationships can predict disease risk when it is familial. Therefore, we developed LRFR, a procedure that is designed to focus entirely on how well a disease is predicted in any individual by his or her genetic relatedness to a diseased family member, with Beta coefficients providing insights to the strength of the association. For any disease D of interest, the LRFR approach proceeds as follows. For every patient A with disease D, an instance is created for each other patient B within A's family. In this instance, the independent predictor variable is the relatedness of A and B, and the binary dependent variable indicates whether B also has the disease D ('label' in Fig. 2). Under this procedure, each pair of patients with disease D in a family gives rise to two identical examples, so we remove one redundant example in every such case.

Besides only considering relatedness of each pair of patients, to reduce confounding we repeated the analysis accounting for four related covariates: differences in gender, age, generation and length of EHR data. Length of EHR data was defined by time, in years, between first and last visit. We provide these created instances as input data for logistic regression and ran a 3-fold cross validation to estimate the Beta coefficients for each disease. We also performed a permutation test, permuting the labels of individuals (for having a specific disease or not) 'across' and 'within' affected families to determine the appropriate significance threshold. As shown in Results, we compare the Beta coefficients and Wald test *P*-values of coefficients for congenital codes against those for a collection of non-congenital codes (accidental falls) via a Mann–Whitney test, testing the null hypothesis that there is no enrichment for congenital codes to have smaller *P*-values or larger Beta coefficients.

2.2.2 LRFR validation

To compare the use of a large collection of high-throughput machine-constructed families against the use of a (necessarily smaller) collection of human-reported families, we also applied the same LRFR methodology to families identified in the Marshfield Clinic Personalized Research Project (PMRP) (McCarty *et al.*, 2005). PMRP is a cohort of over 20 000 adult Marshfield Clinic patients recruited for research. The mean and median age were 58-years, with most subjects having over 30 years of EHR data. During recruitment, patients were asked to self-report all first-degree relatives. Of the 20 000 PMRP participants, nearly 12 000 can be linked to other family members in the Marshfield Clinic EHR totaling 16 400 individuals and 4 515 families.

2.2.3 Disease mapping

For genetic association testing, 4 045 unrelated PMRP individuals with pre-existing Illumina HumanExomeCore BeachChip SNP data were used. These individuals were originally genotyped as part of the AMD Gene Chip Consortium (Fritsche *et al.*, 2015) and are defined as probands for this experiment. Unrelatedness was confirmed by calculating a genetic correlation matrix for all possible pairs; no two pairs had a familial coefficient >0.0884. Four SNPs (rs887829, rs964184, rs4349859 and rs3750847), that are known to be associated with four separate disease phenotypes (hyperbilirubinemia, pure hyperglyceridemia, ankylosing spondylitis and AMD, respectively), were analyzed with and without family data (Supplementary Fig. S2). In this instance, cases were defined by those with the disease specific ICD9 code (ICD9 277.4, 272.1, 270.0 and 362.51, respectively) while controls were defined by those without any related code (i.e. ICD9 277*, 272*, 720* and 362*, respectively) (Ye *et al.*, 2015). Representing a standard case-control study of unrelated individuals, SNP-disease associations were measured in probands only. *P*-values were calculated using a likelihood ratio test based on Firth logistic regression (Firth, 1993), which included age at last visit, length of EHR data and gender as covariates.

We then compared these association results with an analysis using family data. For either self-reported or predicted family members, SNP genotype dosage was imputed given the allele frequency in the population and possible segregation patterns dictated by the genotyped proband. For example, if a proband has a genotype of AA (allele dosage = 2), and A has an allele frequency 12.5% in the population, the imputed allele dosage of the proband's child would be 1.125 assuming Mendel's laws and Hardy Weinberg equilibrium. In rare instances, multiple unrelated probands were genotyped within the same family (e.g. two unrelated parents). In this instance, both genotyped individuals were used to predict allele dosage for family members. Allele dosage (range 0–2) and covariates listed above were input variables into ASSOC under default options for a binary trait and one marker as part of S.A.G.E: version 6.4.1 [2016] (http://darwin.cwru.edu). ASSOC can measure genetic associations from a mixed population of families and singletons. Singletons represent individuals with no family data (i.e. proband only) (Supplementary Fig. S2). A likelihood ratio test was used to calculate *P*-values in the family datasets, consistent with the analysis of unrelateds.

3 Results

3.1 Family pedigree prediction

After applying the family prediction algorithms to 2.6 million patients in Marshfield Clinic's EHR using basic demographic data, we identified 579 561 individuals linked to 173 368 predicted families. In other words, 22% of the total clinic population could be linked to another family member. Of the 173 368 families, 77 519 (45%) families had two or more children identified whereas 141 175 (81%) and 108 859 (63%) families had the mothers and fathers identified, respectively. The number of generations ranged from two to five and the average family size was three (Fig. 3). The largest predicted family contained 33 members, and 9.2% of the predicted families

had over two generations linked to Marshfield Clinic's EHR. Importantly, the cohort of patients in families had on average 10.8 years of medical record data (Table 1).

During manual validation, there were 71 and 188 family members in the randomly selected standard and half-sib families, respectively. All of the predicted standard familial relationships were validated, except that one child was identified as being adopted. In the half-sib families, 96% of the familial relationships were validated; again, one individual was identified as adopted. In the process of manual chart review, 11 and 68 first-degree family members were identified that were not captured electronically in the standard and half-sib families, respectively (Table 2).

3.2 Disease-specific studies

3.2.1 LRFR

Initial disease-specific studies were directed towards color blindness and MD. These two phenotypes were selected because they are well known to be highly heritable, are predominantly monogenic, and have well-defined ICD9 codes. The predominant forms of color blindness and MD are red-green color blindness (deuteranopia) and Duchenne MD, respectively, with pathogenic variants mapped to the X chromosome (X-linked). There were 441 and 194 color blind and MD families totaling 2127 and 688 family members, including 456 and 209 affecteds, respectively. Males were primarily affected in both conditions (Fig. 4) consistent with X-linked disease. There was a fraction of females also affected with MD given there are other forms of MD that map to the autosome, and women who are carriers of pathogenic variants in the Duchenne gene (*DMD*) may also have, albeit less severe, MD. Where multiple family members were coded for either condition, transmission was primarily through the maternal lineage. For example, there were 152 third-generation males identified in 194 MD families; 44 individuals (29%) were coded with MD. Of these 44 individuals, three (6%) had an affected mother and 0% had an affected father. When looking at the second-generation biological uncles in the maternal lineage, 33% of affected uncles had affected mothers. This same pattern was observed for males diagnosed with color blindness, although there were two third-generation females diagnosed with color blindness that had an affected father, suggesting the mother must also have been a carrier. No family had two parents, three generations, or a grandparent and grandchild coded for either phenotype, emphasizing the inherent sparsity of phenotypic data in an EHR even for highly penetrant Mendelian phenotypes.

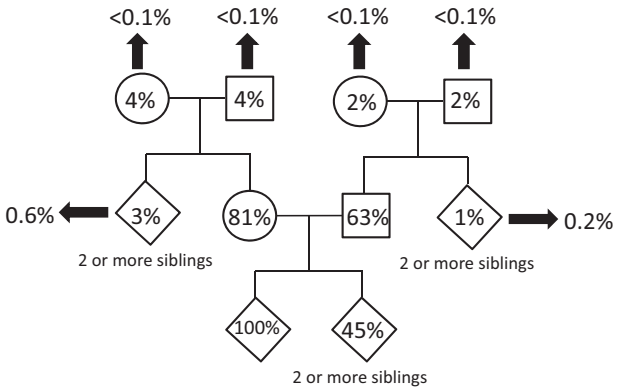


Fig. 3. General pedigree structure. General composition of all families of varying structures in the extended family cohort. Provided in each symbol (□ = males, ○ = females and ◊ = unisex) is the percent of pedigrees with this individual predicted from the EHR. Further extensions of the general pedigree structure are represented by arrows with percentages of families with these extensions provided

To further assess the utility of large populations of families linked to an EHR, we attempted to measure heritability for color blindness and MD. Using a variety of standard methods, heritability for neither disease could be reliably measured. To address this, we developed LRFR to assess the ability of presumed genetic relationships to predict disease concordance among pairs of individuals in de-identified family data with sparse phenotypic information (Fig. 2). In the instance of color blindness and MD, LRFR had strong association results ($P = 2.7E\text{-}6$, Beta coefficient 3.1 and $P = 1.2E\text{-}7$, Beta coefficient 3.7, respectively).

To evaluate LRFR's capacity to segregate diseases expected to have heritable influences from those without, we focused on 28 additional phenotypes. The diseases included 19 'congenital' phenotypes

Table 1. Demographics of the large family cohort

Number of generations	Number of families	Number of patients	Family size (mean)	Years of EHR data (mean)
2	161 489	507 463	3.1	10.50
≥3	11 879	72 098	6.1	12.86
Total	173 368	579 561	3.3	10.81

Table 2. Manual chart review of 40 randomly selected standard and half-sib families

Type	Generation	Number individuals	Male (FP)	Female (FP)	Male (FN)	Female (FN)
Standard	Grandparent	0				
	Parent	32	0 (0.0%)	0 (0.0%)	2 (6.3%)	1 (3.1%)
	Child	39[a]	0 (0.0%)	0 (0.0%)	1 (2.6%)	7 (17.9%)
	Total	71	0 (0.0%)	0 (0.0%)	3 (4.2%)	8 (11.3%)
Half-Sib	Grandparent	69	0 (0.0%)	0 (0.0%)	2 (2.9%)	2 (2.9%)
	Parent	100[a]	6 (6.0%)	1 (1.0%)	33 (33.0%)	18 (18.0%)
	Child	19	1 (5.3%)	0 (0.0%)	4 (21.1%)	9 (47.4%)
	Total	188	7 (3.7%)	1 (0.5%)	39 (20.7%)	29 (15.4%)

FP, false positive; FN, false negative.

[a]One individual was identified as being adopted.

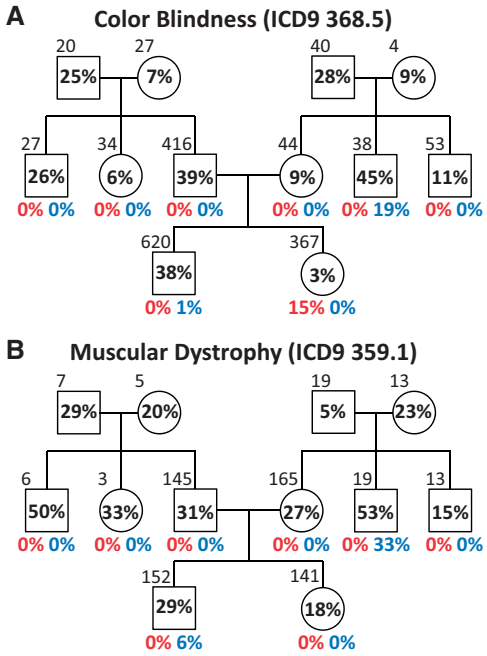


Fig. 4. Summary of all families diagnosed with color blindness or muscular dystrophy. A representation of all families of varying structures diagnosed with either color blindness (**A**) or muscular dystrophy (**B**). Above each symbol (□ = males and ○ = females) are the total number of individuals identified. Inside each symbol is the percent affected. Below each symbol is the percent effected that have either an affected father (red) or mother (blue). For example, there were 620 identified third generation males across all color blind families, 38% were affected and 1% of the affected third generation males had an affected mother

(ICD9 741–759) with likely genetic etiologies. For example, polycystic kidney disease, fragile X syndrome and Marfan's disease are broadly captured by these codes. Nine additional phenotypes that consisted of accidental falls (ICD9 E880-E888) were selected as control phenotypes. For all phenotypes, disease prevalence ranged from 0.20 to 0.28 in affected families. There was little to no correlation in disease status of affected families for any two ICD9 codes. The mean, median and standard deviation of correlation coefficients for all possible pairs of 28 ICD9 codes were −0.013, −0.010 and 0.032, respectively.

When applying LRFR to the 28 disease phenotypes, disease status was statistically associated with genetic relatedness for 15 phenotypes after adjustment for multiple hypothesis testing ($P < 0.0018$ assuming 28 tests and $\alpha < 0.05$). The most significant associations included ICD9 code 755 defining 'Other congenital anomalies of limb' (Beta coefficient 2.2, $P = 2.1E\text{-}57$), ICD9 code 747 defining 'Other congenital anomalies of circulatory system' (Beta coefficient 3.2, $P = 2.5E\text{-}24$), and ICD9 code 752 defining 'Congenital anomalies of genital organs' (Beta coefficient 2.3, $P < 1.2E\text{-}26$) (Table 3). Importantly, there was an enrichment for congenital phenotypes with significant associations as ranked by LRFR *P*-values (Mann–Whitney $P = 0.034$). In addition, the strength of associations, as defined by Beta coefficients, were also larger in congenital phenotypes compared to accidental falls (Mann–Whitney $P = 0.023$).

On adjusting for covariates that may influence disease status (age, sex, length of longitudinal data and generational differences), the top associations continued to be the same congenital codes. Of the 15 that were originally statistically significant when covariates were not included in the model, four were no longer significant, including two E-codes (Table 3). There was also a strong association between disease status and genetic relatedness for E-code E884 in both analyses. This may emphasize confounding effects where environmental/social influences may be stronger for those who are more genetically related. On applying the Mann–Whitney test to ranked Beta coefficients and *P*-values when covariates were included, the associations from *P*-values were borderline significant (Mann–Whitney $P = 0.061$) whereas there was an improvement when ranked by Beta coefficients (Mann–Whitney $P = 0.0051$) (Table 3). For additional disease specificity, full LRFR results for all non-rare (>9 cases) congenital sub-codes are available in Supplementary Table S1.

In permuting the labels of individuals (for having a specific disease or not) 'across' affected families to determine the significance threshold, congenital phenotypes remained significant, with most permuted phenotypes having Beta coefficients approaching zero. Because the diagnosis of some diseases may be influenced by social factors within families, labels were permuted again, but only 'within' affected families. In this case, Beta coefficients for conditions with significant LRFR results again remained significantly higher than the permuted results. Supplementary Table S2 presents empirical *P*-values from this permutation test, where the *P*-value for each disease is the fraction of Beta coefficients from 10 000 permutations that are higher than the value reported for that disease in Table 3. These results demonstrate that LRFR is not likely biased by family structures, disease prevalence, or familial influences pertaining to seeking healthcare.

3.2.2 LRFR validation

To illustrate the value of the automatically constructing pedigrees, we repeated the LRFR analysis in a subset of Marshfield Clinic patients who were recruited as part of Personalized Medicine Research Project (PMRP) with self-reported familial relationships (McCarty *et al.*, 2005). The number of individuals in this dataset was much smaller and older than in the previous dataset; but these patients often captured older generations and had more complete longitudinal data, potentially improving phenotypic ascertainment. The median disease frequency for the 28 phenotypes in PMRP-linked families was 2.8-fold higher compared to the larger family cohort. On analyzing the presumed heritable and control phenotypes, again the top associations included congenital phenotypes (Supplementary Table S3), but the associations did not segregate with disease type when ranked by Beta coefficients or *P*-values regardless of the use of covariates (Mann–Whitney $P > 0.73$). These results provide evidence for the benefit of large, automatically constructed families for pedigree analysis in EHRs, compared to use of the smaller number of available self-reported families.

3.2.3 Disease mapping

Under certain circumstances family studies often have higher statistical power for genetic mapping (Gray-McGuire *et al.*, 2009), but as previously mentioned, identifying, collecting and genotyping family members can be costly and time consuming. To evaluate the power of automatically collected families in an EHR for disease mapping, we compared association test results with and without family data (Supplementary Fig. S2). Of the 4045 unrelated individuals in PMRP with genetic data, and depending on disease, ∼1900 individuals had predicted family data; this is compared to nearly 1400 individuals with self-reported family data. Focus was towards four SNPs well known to be associated with four separate disease phenotypes including rs3750847 (*ARMS2*), rs887829 (*UGT1A1*), rs964184 (*ZPR1*) and rs4349859 (*HLA-B27*) that are associated with age

Table 3. LRFR association results for 28 phenotypes from the large family cohort

ICD9	Description	Affecteds	Affected families	Without covariates		With covariates	
				Beta coefficient	*P*-value	Beta coefficient	*P*-value
Congenital phenotypes							
741	Spina bifida	535	510	−0.26	0.78	0.26	0.85
742	Other congenital anomalies of nervous system	2385	2276	0.40	0.32	0.082	0.88
743	Congenital anomalies of eye	9548	8347	0.72	8.8E-10	−0.19	0.24
744	Congenital anomalies of ear face and neck	1509	1481	3.0	2.0E-04	2.1	0.12
745	Bulbus cordis anomalies and anomalies of cardiac septal closure	5100	4736	2.2	4.1E-18	1.5	1.2E-4
746	Other congenital anomalies of heart	4141	3958	1.9	4.3E-10	1.9	8.0E-06
747	Other congenital anomalies of circulatory system	4251	4049	3.2	2.5E-24	2.3	6.0E-07
748	Congenital anomalies of respiratory system	1085	1064	2.8	0.0031	2.9	0.063
749	Cleft palate and cleft lip	761	733	2.1	0.0040	1.9	0.12
750	Other congenital anomalies of upper alimentary tract	3639	3431	3.8	1.4E-32	3.4	6.7E-11
751	Other congenital anomalies of digestive system	1647	1612	3.0	7.0E-06	3.8	5.0E-05
752	Congenital anomalies of genital organs	7333	6952	2.3	1.2E-26	1.8	3.1E-07
753	Congenital anomalies of urinary system	3511	3353	0.59	0.10	0.65	0.16
754	Certain congenital musculoskeletal deformities	14 360	12 827	1.1	1.1E-26	0.86	1.96E-08
755	Other congenital anomalies of limbs	11 120	10 235	2.2	2.1E-57	1.9	3.9E-18
756	Other congenital musculoskeletal anomalies	7088	6634	1.4	4.1E-13	0.93	6.5E-4
757	Congenital anomalies of the integument	13 939	12 529	1.1	1.7E-24	1.1	7.1E-12
758	Chromosomal anomalies	1773	1692	1.1	0.012	1.2	0.058
759	Other and unspecified congenital anomalies	178	176	3.5	0.24	3.7	0.31
Accidental falls							
E880	Accidental fall on or from stairs or steps	5756	5396	−0.036	0.89	−0.81	0.011
E881	Accidental fall on or from ladders or scaffolding	1423	1413	0.15	0.95	−3.8	0.10
E882	Accidental fall from or out of building or other structure	583	580	1.7	0.50	1.0	0.80
E883	Accidental fall into hole or other opening in surface	414	412	−60	0.99	−34	0.99
E884	Other accidental falls from one level to another	11 111	9945	1.5	1.5E-36	0.89	5.3E-07
E885	Accidental fall on same level from slipping tripping or stumbling	15 496	13 570	0.52	3.1E-07	0.16	0.23
E886	Fall on same level from collision, pushing, or shoving, by or with other person	2738	2645	2.0	2.7E-06	−0.38	0.48
E887	Fracture, cause unspecified	574	567	1.5	0.33	2.5	0.41
E888	Other and unspecified fall	8332	7618	0.42	0.020	−0.38	0.076

related macular degeneration risk, bilirubin metabolism, cholesterol metabolism and ankylosing spondylitis risk, respectively.

In the case-control study design of unrelated individuals, all four SNPs were associated with their respective phenotype ($P \leq 5.5$E-5). For example, there were 54 and 3863 unrelated cases and controls for hyperbilirubinemia, respectively, that were associated with rs887829 genotype ($P = 9.6$E-34). Rs887829 tags for the *UGT1A1*28* allele associated with loss of UGT1A1 enzyme function (*Iyer et al.*, 2002). By incorporating self-reported or predicted family members, association results became stronger as indicated by dramatically smaller *P*-values ($P = 1.3$E-178 and $P = 3.6$E-143, respectively; Table 4). In total, incorporating predicted family data improved association results for all four SNP-disease pairs compared to the analysis of unrelated individuals. When comparing association results between the different family types, self-reported families had comparable but smaller *P*-values than predicted families for all SNP-disease pairs. Regardless, whereas using additional family data never materially reduced the power, these results suggest that incorporating additional family data readily attainable in an EHR, even if not directly genotyped, may improve statistical power for genetic mapping.

4 Discussion

In this study, we were able to link 22% of Marshfield Clinic's current and historical patient population to another family member with high accuracy using standard but de-identified demographic data available in an EHR, including half-sib families. Identifying families in an EHR can provide a highly valuable resource for selecting subjects for enrollment into family-data-collection studies, short-circuiting a huge amount of effort and cost in the process.

Although we provide evidence that EHR-linked families may have great value in research, this method does have limitations that are influenced by inherent attributes of the EHR. The ability to identify families will greatly depend on the quality and longitudinal nature of data within an EHR. With Marshfield Clinic's EHR dating back to 1984, identifying generations that left the household prior 1984 may be difficult. This will be further complicated by patients who geographically move in and out of a healthcare system. These limitations are exemplified by predicted pedigrees representing predominantly small nuclear families (Table 1 and Fig. 3) and high false negative rates observed during manual assessment of predicted families (Table 2). These temporal limitations may also influence the types of phenotypes that can be studied. If most families represent the youngest generations, studying age dependent diseases may be challenging. This may explain the difference in genetic association results for age related macular degeneration where the smaller but older self-reported families had stronger association results compared to the younger but larger sample of predicted families (Table 4). To improve the prediction algorithm, future research may leverage public records such as birth records to reduce the false negative rate and capture older generations.

Another limitation of this study is rooted in the phenotypes that were extracted from the EHR (i.e. ICD9 codes). Although ICD9

Table 4. Genetic association results showing value of family data for disease mapping

SNP	Gene	Disease	ICD9	Case-control of unrelateds		Self-reported families			Predicted families		
				A/U	*P*-value	A/U families	A/U singletons	*P*-value	A/U families	A/U singletons	*P*-value
rs887829	*UGT1A1*	Hyperbilirubinemia	277.4	54/3863	9.6E-34	62/1153	34/2583	1.3E-178	55/1248	34/2484	3.9E-143
rs3750847	*ARMS2*	AMD	362.51	482/2363	5.5E-5	406/809	381/1495	5.7E-12	145/1155	411/1391	1.3E-6
rs964184	*ZPR1*	Pure hyperglyceridemia	272.1	300/676	2.1E-6	376/839	195/434	4.2E-13	309/991	195/339	8.3E-9
rs4349859	*HLA-B27*	Ankylosing spondylitis	720.0	102/3943	6.7E-8	95/1119	70/2634	5.1E-37	75/1223	69/2530	1.7E-36

A, affected; U, unaffected; AMD, age-related macular degeneration.

coding is frequently used to identify cases and controls for genetic research (Ye *et al.*, 2015; Hebbring *et al.*, 2015; Rastegar-Mojarad *et al.*, 2015), it should be mentioned that ICD9 coding, and now ICD10, is applied in the United States primarily for billing. ICD coding can change over time and can be employed differently between physicians and across healthcare institutions (Hebbring, 2014). It may be expected that these limitations are temporary as EHRs become standardized, longitudinally mature and more portable across healthcare systems.

In addition to the development of a family prediction algorithm, we further provide flexible methods that leverage EHR-linked families for genetic epidemiologic research. Even with the limitations described above, LRFR results demonstrate that individuals coded for congenital phenotypes were enriched in families and were correlated with degree of relatedness (Tables 3 and 4). It is conceivable that methods like LRFR may help researchers identify the most interesting families and diseases for future genomic research. When limited genetic data are readily available, we further demonstrate that even sparse family data may improve genetic association testing without the need of additional costs associated with recruitment and genotyping of family members (Table 5). Future studies are warranted to better understand how different family structures, population sizes, genetic effect sizes, minor allele frequencies and disease models may influence genetic association testing using predicted families. These results may be highly relevant as EHR-linked biobanks continue to grow (Kho *et al.*, 2011; McCarty *et al.*, 2011; Gottesman *et al.*, 2013; Krishnamoorthy *et al.*, 2014), including efforts to recruit over 1 million United States residences linked to EHR data through the 'All of Us Research Program' (formally known as Precision Medicine Initiative) (Collins and Varmus, 2015).

The implications from this work may have broad clinical applications, including better estimation of the familial component of risk across a wide range of diseases and more accurate predictive models for future health trajectories and treatment responses for patients. One exciting future research direction would be to incorporate this newly constructed pedigree information into disease risk models that use patient clinical history in the EHR, including laboratory results, prescriptions, diagnoses, procedures and text-extracted signs and symptoms.

Family histories have long been a powerful tool in primary care, with well-documented clinical validity and utility (Rich *et al.*, 2004; Rubinstein *et al.*, 2011; Qureshi *et al.*, 2012; USPSTF, 2014). Interview- or survey-based patient family histories obtain information on family structure, patient age, gender, ethnicity and disease history for several conditions. Although such family histories are highly useful, they are often limited by a patient's memory, awareness, understanding and willingness to share (Ashida and Schafer, 2015). Likewise, family histories may not be updated frequently and are restricted to only a few diseases. Identifying clinical phenotypes with a strong familial component may provide a foundation for future clinical decision support tools designed to utilize, in real time, patient and family medical records to request or manage personalized disease-specific family histories. These potential clinical tools may be further complemented by advancements in genomic medicine where large patient populations may have extensive genetic data available. Although we have established potential uses of EHR-linked families for genetic research, it is expected that many other phenotypes can be studied in EHR-linked families, including infectious diseases that are transmitted in families independent of genetic relatedness.

Before EHR-linked family data is widely accepted in research or clinical care, there are ethical considerations. If applied to clinical care, some patients may not want to share or receive family data. In research, there should be caution when utilizing genetic and phenotypic data from family members who have not consented for research. In the instance of this study, we took great care to ensure that family, genetic and phenotypic data could not be directly mapped back to identifiable information.

In conclusion, this study demonstrates that an EHR system can efficiently and effectively produce family pedigree data that can be used for genetic epidemiologic research. Furthermore, this study may provide an intriguing perspective for the future of precision medicine, specifically, the future when large patient populations in defined families are unified with genomic data in an integrated EHR system.

Acknowledgements

The authors would like to thank Drs. Steven Schrodi and Yeunjoo Song for their consultations, Rachel Stankowski and Marie Fleisner for their editing assistance and the Reviewers for their insightful and helpful comments.

Funding

This work was supported by NIH grants: NCATS grant 9U54TR000021, NCRR grant 1UL1RR025011, NLM grant 1K22LM011938 and 5T15LM007359, NHGRI 1U01H6006389 and NIGMS 1R01GM114128. The authors also gratefully acknowledge the support from Marshfield Clinic Research Foundation and its generous donors.

Conflict of Interest: none declared.

References

Ashida,S., and Schafer,E.J. (2015) Family health information sharing among older adults: reaching more family members. *J. Commun. Genet.*, **6**, 17–27.

Chong,J.X. *et al.* (2015) The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. *J. Hum. Genet.*, **97**, 199–215.

Collins,F.S., and Varmus,H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **372**, 793–795.

Emilsson,L. *et al*. (2015) Autoimmune disease in first-degree relatives and spouses of individuals with celiac disease. *Clin. Gastroenterol. Hepatol*., **13**, 1271–1277.

Firth,D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Fritsche,L.G. *et al*. (2015) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet*, **48**, 134–143.

Gottesman,O. *et al*. (2013) The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med*., **15**, 761–771.

Gray-McGuire,C., Bochud,M., and Goodloe,R. (2009) Genetic association tests: A method for the joint analysis of family and case-control data. *Hum. Genomics*, **4**, 2–20.

Hagberg,A.A. *et al*. (2008) Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux, G. *et al.* (ed.) *Proceedings of the 7th Python in Science Conference* (SciPy2008), Pasadena, CA, pp. 11–15.

Hebbring,S.J. (2014) The challenges, advantages and future of phenome-wide association studies. *Immunology*, **141**, 157–165.

Hebbring,S.J. *et al*. (2015) Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinformatics*, **31**, 1981–1987.

Iyer,L.S. *et al*. (2002) UGT1A1*28 polymorphism as a determinant of irinotecan disposition and toxicity. *Pharmacogenomics J*., **2**, 43–47.

Kho,A.N. *et al*. (2011) Electronic medical records for genetic research: results of the eMERGE consortium. *Sci. Transl. Med*., **3**, 79re1.

Krishnamoorthy,P. *et al*. (2014) A review of the role of electronic health record in genomic research. *J. Cardiovasc. Transl. Res*., **7**, 692–700.

Manolio,T.A. *et al*. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Mayer,J. *et al*. (2014) Use of an electronic medical record to create the marshfield clinic twin/multiple birth cohort. *Genet. Epidemiol*., **38**, 692–698.

McCarthy,M.I. *et al*. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet*., **9**, 356–369.

McCarty,C.A. *et al*. (2011) The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics*, **4**, 13.

McCarty,C.A. *et al*. (2005) Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Pers. Med*., **2**, 9–79.

Qureshi,N. *et al*. (2012) Effect of adding systematic family history enquiry to cardiovascular disease risk assessment in primary care: a matched-pair, cluster randomized trial. *Ann. Intern. Med*., **156**, 253–262.

Rastegar-Mojarad,M. *et al*. (2015) Opportunities for drug repositioning from phenome-wide association studies. *Nat. Biotechnol*., **33**, 342–345.

Rich,E.C. *et al*. (2004) Reconsidering the family history in primary care. *J. Gen. Intern. Med*., **19**, 273–280.

Rubinstein,W.S. *et al*. (2011) Clinical utility of family history for cancer screening and referral in primary care: a report from the Family Healthware Impact Trial. *Genet. Med*., **13**, 956–965.

U.S. Preventive Services Task Force (USPSTF). (2014) *Guide to Clinical Preventive Services*. Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services.

Visscher,P.M. *et al*. (2012) Five years of GWAS discovery. *Am. J. Hum. Genet*., **90**, 7–24.

Welter,D. *et al*. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*., **42**, D1001–D1006.

Wright,S. (1922) Coefficients of inbreeding and relationship. *Am. Nat*., **56**, 330–338.

Ye,Z. *et al*. (2015) Phenome-wide association studies (PheWASs) for functional variants. *Eur. J. Hum. Genet*., **23**, 523–25529.