Estimate of disease heritability using 4.7 million familial relationships inferred from electronic health records

Fernanda Polubriaginof[1], Kayla Quinnies[1,2*], Rami Vanguri[1*], Alexandre Yahi[1], Mary Simmerling[3], Iuliana Ionita-Laza[4], Hojjat Salmasian[1,5], Suzanne Bakken[1,6], George Hripcsak[1], David Goldstein[2], Krzysztof Kiryluk[7], David K. Vawdrey[1,5 †], Nicholas P. Tatonetti[1,2,7,8,†]

[1] Department of Biomedical Informatics, Columbia University, New York, NY
[2] Institute for Genomic Medicine, Columbia University, New York, NY
[3] Department of Medicine, Weill Cornell Medicine, Cornell University, New York, NY
[4] Mailman School of Public Health, Columbia University, New York, NY
[5] Value Institute, NewYork-Presbyterian Hospital, New York, NY
[6] School of Nursing, Columbia University, New York, NY
[7] Department of Medicine, Columbia University, New York, NY
[8] Department of Systems Biology, Columbia University, New York, NY
[*] These authors contributed equally, ordered alphabetically
[†] Co-senior author

Correspondence:
Nicholas P. Tatonetti, PhD
622 W 168th. St. PH20-402
New York, NY 10032
USA
nick.tatonetti@columbia.edu

Abstract

Heritability is a fundamental characteristic of human disease essential to the development of a biological understanding of the causes of disease. Traditionally, heritability studies are a laborious process of patient recruitment and phenotype ascertainment. Electronic health records (EHR) passively capture a wide range and depth of clinically relevant data and represent a novel resource for studying heritability of many traits and conditions that are not typically accessible. In addition to a wealth of disease phenotypes, nearly every hospital collects and stores next-of-kin information on the emergency contact forms when a patient is admitted. Until now, these data have gone completely unused for research purposes. We introduce a novel algorithm to infer familial relationships using emergency contact information while maintaining privacy. Here we show that EHR data yield accurate estimates of heritability across all available phenotypes using millions familial relationships mined from emergency contact data at two large academic medical centers. Estimates of heritability were consistent between sites and with previously reported estimates. Inconsistencies were indicative of limitations and opportunities unique to EHR research. Critically, these analyses provide a novel validation of the utility of electronic health records in inferences about the biological basis of disease.

Introduction

Family history is one of the most important disease risk factors necessary for the implementation of precision medicine in the clinical setting[1]. The predictive value of family history for any given trait is directly related to the fraction of phenotypic variance attributable to genetic factors, known as heritability[2]. Knowledge of disease heritability combined with family history information is clinically useful for identifying risk factors, estimating risk of disease, customizing treatment, and tailoring patient care. Moreover, by quantifying genetic contribution to a trait, heritability estimation represents the first step in gene mapping efforts for any disease.

Estimating heritability has traditionally required in-depth family studies, with twin studies being the gold standard. By their nature these studies can be laborious, limiting their sample sizes and, subsequently, their power. A notable exception, and perhaps the largest single study, used 80,309 monozygotic and 123,382 same-sex dizygotic twins to conclude that there is significant familial risk for prostate, melanoma, breast, ovary, and uterine cancers[3]. Another study brought together 2,748 twin studies conducted since 1955 covering 14.5 million subjects. However, in such a meta-analysis individual data are not available, preventing any study of cross-sections, combinations of traits, or strata that were not analyzed in the original study[4].

Electronic Health Records (EHR) are in broad use and offer an alternative to traditional phenotyping. Everyday, the EHR records thousands of patient phenotypes from drug prescriptions and disease diagnoses to clinical pathology results and physician notes. Use of the EHR as an observational dataset presents a novel opportunity to conduct rapid and expansive studies of disease and phenotype heritability. In particular, they enable access to traits that otherwise might not be studied. In addition, data captured by these systems represent the diversity of the patient populations they serve, and, in ethnically diverse regions like New York City, make previously unattainable cohorts available for study.[5] The caveat is that these data are known to contain many biases and errors that limit their use. Issues regarding missingness and accuracy are widely cited as the primary limitations[6]. However, the most critical limitation for genetic studies may be the uncontrolled ascertainment bias. The probability that a particular trait is recorded in the EHR is not uniform across disease conditions or across patients. For example, a patient seen for a routine checkup with no symptoms is unlikely to undergo an MRI, regardless of whether or not they have an unruptured brain aneurysm.

The genetic relatedness between patients is not routinely captured in the EHR during clinical practice. In some hospitals, as is the case for the two we represent, a link is made between the mother's and child's medical records upon birth. In general, however, familial links are not present. Recent work has identified twins by comparing birth dates and surnames[7], but there is a more comprehensive source of familial relationship data that is available at nearly every hospital across the country – the emergency contact information. Upon admission, each patient is asked to provide contact details to be used in case of emergency as well as how they are related to the individual provided. If accurate, this ubiquitous resource can be used to define a broad network of relatedness across a hospital's patient population.

In this study, we demonstrate the utility of the EHR as a genetics research resource by using extracted data to estimate the heritability and familial recurrence rates of over 700 phenotypes -- both quantitative and dichotomous. We performed this analysis independently at two large academic medical centers in New York City. We present our algorithm for extracting relationships, called Relationship Inference From The Electronic Health Record (RIFTEHR), and use it to infer 4.7 million familial relationships among our patients. We then computed recurrence rates and heritability estimates for every available phenotype. Our derived heritability estimates accurately reflect those previously reported and we report heritability estimates for many traits that may otherwise never have been studied.

Results

Mining familial relationships from the EHR

We obtained the data for this study from the inpatient EHR used at the hospitals of Columbia University Medical Center and Weill Cornell Medical College. These hospitals operate together as NewYork-Presbyterian Hospital and herein, we will refer to the hospitals and the data associated with them as Columbia and Cornell, respectively. The study was approved by Institutional Review Boards at both Columbia and Cornell University.

In total, 4,768,013 emergency contacts were provided by 2,388,455 patients at the two medical centers. Of these, we identified the emergency contact as a patient in 785,943 cases (488,932 and 297,011 at Columbia and Cornell, respectively). Using these next-of-kin data, we inferred an additional 2,614,657 relationships at Columbia and 1,200,977 at Cornell. Including inferences, a total of 3,103,589 unique relationships have been identified at Columbia and 1,497,988 at Cornell. Inferred relationships include first to fourth degree relatives as well as spouses and in-laws (Table 1, Supplementary Table 1). We grouped individuals into families by identifying disconnected subgraphs (*Materials and Methods*). We found 223,307 families at Columbia containing 2 to 134 members per family. Similarly, we found 155,883 families at Cornell, with up to 129 members per family. This includes 127 families that span four generations (i.e. families that contain great-great-grandparents and great-great-grandchildren) at Columbia and 72 families that span four generations at Cornell.

The relationship between mother and child was explicitly documented in the EHR for babies delivered at both medical centers. This 'EHR mother-baby linkage' provided a reference standard for maternal relationships, allowing us to compute sensitivity and positive predictive value (PPV) of the relationship inference method. For maternal relationships, we obtained 92.9% sensitivity with 95.7% PPV at Columbia and 96.8% sensitivity with 98.3% PPV at Cornell (Figure 1A).

We validated the identified relationships by comparison to genetic relatedness (Figure 1). We collected a dataset of 186 patients for which we have EHR-inferred relationships and who have genetic data available that was consented for reuse. We used PLINK to estimate relatedness. All 78 predicted parent/child relationships had the expected genetic relatedness of 50% as well as the three grandparent/grandchild relationships. All 19 sibling relationships were genetically related, but four were identical twins and two were half-siblings. Overall, relationships extracted from the EHR significantly correlate with the expected genetic relatedness ($r = 0.65$, $p = 6.26e\text{-}14$, Figure 1B).

Health records-based estimates of heritability

To differentiate heritability estimates derived under uncertain ascertainment conditions, we introduce the concept of "observational h^2" or h_2^o. h_2^o is an estimate of the narrow-sense heritability where the phenotypes (traits) come from observational data sources. Observational data are subject to confounding biases from physician and patient behaviors that will affect the probability that a particular trait is ascertained. These ascertainment biases can vary from patient to patient, family to family, and cases to controls. The consequence is that the estimated heritability will be highly dependent on the particular families and individuals upon which the estimate is based. To correct for this, we bootstrapped the heritability estimates. For each sampling we used SOLAR[8] to estimate the heritability of the trait, in a procedure we call SOLARStrap (*Materials and Methods*). High sampling variance indicates the presence of heterogeneous biases. Heritability estimates are adjusted for age and sex.

We mined the literature for heritability estimates and found 91 phenotypes that mapped to phenotypes we curated from the EHR. We used the Columbia data to set the quality control parameters of the SOLARStrap procedure (*Materials and Methods*). 10 of the traits in the Cornell data passed these QC criteria and we found that they were significantly correlated with literature estimates for these traits ($r=0.73$, $p=0.016$, Figure 2A). On average, estimates from Columbia were 20% ± 9% lower than those reported in the literature and those from Cornell were 7% ± 9% lower (Figure 2B). Heritability estimates derived from Cornell data were highly

correlated with those derived from Columbia data (r=0.67, p=2.56e-12, Figure 2C). As a group, respiratory diseases had the highest average heritability for both dichotomous (Figure 2D) and quantitative (Figure 2E) traits. Genitourinary and gastrointestinal diseases had the lowest average heritability.

For dichotomous traits, we explored the relative contribution of genetics and the environment to the phenotype by comparing heritability estimates to sibling recurrence rates (Figure 2F). Disease groups fell into four distinct regions: (1) those with greater than average genetic and environmental contribution (Figure 2F, top right) – respiratory and neurologic diseases fell into this quadrant; (2) those with high genetics and lower than average environment (Figure 2F, top left) – e.g. endocrine and metabolic diseases; (3) high environment and low genetics (Figure 2F, bottom right) – gastrointestinal and genitourinary diseases; and (4) low environment and low genetics (Figure 2F, bottom left) -- the general category of signs and symptoms is an example here.

Using phenotypes from the EHR for heritability can provide clarity for poorly studied traits, reveal subtle differences between closely related conditions, and open up new avenues of heritability research. For example, two previous studies have shown conflicting evidence for the relative heritability of HDL cholesterol and LDL cholesterol [9,10]. The larger of these two studies (N=378) found no difference in the heritability of these two traits when adjusting for age and sex, while the other found a slightly higher heritability for HDL, but was underpowered to detect significance. We present strong evidence that HDL is significantly more heritable than LDL (h_2^o=0.49 vs 0.36, p=5.3e-41 at Columbia; h_2^o=0.47 vs 0.25, p=6.2e-159 at Cornell; Figure 2G). At 96,241 patients in the Columbia cohort and 33,239 patients in the Cornell cohort, ours may be the largest heritability study of cholesterol ever conducted. In addition, subtle phenotypical variations that are routinely collected clinically can be studied. For example, we found that the heritability of “obesity” is significantly greater than for “morbid obesity” (h_2^o =0.43 vs 0.36, p=2.1e-8, N=26,783 at Columbia and h_2^o =0.63 vs 0.51, p=3.1e-9, N=11,220 at Cornell). Finally, the EHR can identify novel traits for genetic study. The most heritable trait we found was for “victim of child abuse,” h_2^o=0.90 (0.73-1.00), N=1,142 (Table 2). This trait is unique in that it is not a trait of the individual with the diagnosis code, but of another individual with whom the child interacts. To account for a potential artifact introduced by several siblings abused by a single individual, we recomputed heritability excluding siblings (*Materials and Methods*). We found that, while the effect is mitigated, the heritability remains high at h_2^o=0.80 (0.68-0.96) (Table S5). The familial trend of this behavioral trait has been well documented in the psychology literature[11-13]. Our findings provide additional evidence for a genetic role as well. Scientists studying child abuse and related conditions may consider performing a more traditional genetics analysis in the future.

Recurrence Rates

We estimate sibling and familial recurrence for 765 dichotomous traits at Columbia and 393 traits at Cornell. When looking at sibling and familial recurrence, perinatal conditions are the most concordant between sites (r^2 = 0.94 for sibling and 0.96 for familial). The least concordant were diseases of the digestive system (r^2 = 0.02) for sibling and signs and symptoms for familial (r^2 = 0). Sibling recurrence and familial recurrence are highly correlated (r = 0.71, p = 2.52e-40) as well as familial recurrence (r = 0.49, p = 1.99e-21) (Figure 3A and 3B). Sibling recurrence, on average, is greater than familial recurrence at both sites (Figure 3C). We also calculated recurrence by disease site and stratified by relationship type (sibling, cousin, first cousin once removed). We observe that disease recurrence among siblings is higher than among cousins, which is higher than among first cousin once removed (Figure 3D).

Data accuracy and missingness

We evaluated the effect of the two most commonly cited limitations of EHR data, errors and missingness, on our estimates of observational heritability (Figure S1). Rhinitis is highly heritable in family studies (h_2=0.95 CI=0.78-0.97) [14] and also has high observational heritability at both sites (h_2^o=0.62 CI:0.49-0.73, h_2^o =0.78 CI:0.61-0.91, Figure 3B). We evaluated the effect of errors and missingness on h_2^o for rhinitis at Cornell. The estimates are robust to missingness (Figure S1B). When 30% of the data are removed, the estimates remain consistent. Note, that as more data are missing, power will become the major limitation. Heritability estimates are consistent until 20%, or more, of the data are noise, at which point the confidence intervals no longer overlap (Figure S1A). Injection of 5% noise reduces the estimate 13% (from h_2^o =0.77 to h_2^o =0.67) and 10% noise reduces the estimate 30% (from h_2^o =0.77 to h_2^o =0.53). This likely explains why our estimates are 7-20% lower than what would be expected from a carefully ascertained study, corresponding to around 5% misclassification in our EHR.

Discussion

Analysis of EHR data has yielded insight into drug effectiveness and allowed precise definition of phenotypes to investigate disease processes [15-20]. For the first time on a large scale, we have used EHR data to infer pedigrees from patient-provided emergency contact information. We present our novel algorithm for performing this relationship extraction, RIFTEHR, and validated its performance. This approach has significant implications for estimating heritability of disease without direct genetic testing. The EHR data used in this research are nearly ubiquitous and, if privacy is adequately protected, could allow almost any research hospital to identify related patients with high specificity and sensitivity. Finally, we used EHR-inferred relationships to evaluate the heritability of 2,089 traits and found 328 with significant heritability. The heritability of many of these traits have never before been studied.

Heritability is a key component in precision medicine, and is typically estimated based on family history. Collection of comprehensive and accurate family history is time-consuming and does not occur during the vast majority of clinical encounters. The construction of pedigrees by inference of relatedness from administrative records allows for rapid assessment of family history and heritability at scales that were previously impossible to achieve. The algorithm used in this study uncovered over 379,000 pedigrees within the medical records of two academic medical centers. We validated the inferred familial relationships against both clinical and genetic references and found PPV between 87% and 99%. One of the limitations of our method is the challenge to differentiate between direct blood relatives and adopted families. Emergency contact is not a biological construct; therefore, patients identify not only direct-blood relatives, but also adoptive family members and use familial labels for friends.

Using EHR-inferred relationships we calculated heritability, sibling recurrence, and familial recurrence estimates among individuals with defined relationships. Previous research in this area has focused on family studies of known relatives, specifically twins. Mayer and colleagues used EHR data to create a cohort of 2,000 twins/multiple births and measured concordance among identified twins for two highly heritable diseases, muscular dystrophy and fragile-X syndrome.[7] Our study looked not only at twins, but entire families across several generations. We evaluated 2,089 traits and computed high confidence heritability estimates for 328 of them. Importantly, most previous studies have predominantly involved White Europeans and may not be representative of other populations. However, our results reflect the diverse, multiethnic population of New York City.

The primary and most significant challenge when using traits defined from an observational resource, like the electronic health records (EHR), is incomplete phenotype information resulting in ascertainment bias. In a heritability study, the phenotype of each study participant is, ideally, carefully evaluated and quantified. This is infeasible, however, when the cohort contains millions of patients with thousands of phenotypes. The differential probability that a given individual will be phenotyped for a study trait is the *ascertainment bias*. The bias may depend on many latent

factors, including the trait being studied, the trait status of relatives, the proximity to the hospital, and an individual's ethnicity and cultural identification, among others. The consequence of this uncontrolled ascertainment bias is that heritability estimates will be highly dependent on the particular individuals in the study cohort. We used repeated sub-sampling to characterize this dependence quantitatively. EHR-based heritability estimates are particularly well-suited for complex traits that require large numbers of patients (e.g., Type 2 Diabetes Mellitus and Obesity). Most importantly, using the EHR can identify new avenues for research. We report very high heritability for child abuse, indicating a potential genetic role in this well studied condition.

The unique nature of the relationships and phenotypes derived from the EHR may necessitate novel methods for estimating heritability. We used a mixed linear model implemented in SOLAR[8] to estimate heritability and used repeated sampling to characterize the variance from ascertainment heterogeneities. There may be more accurate ways to estimate heritability from this unique data source. For example, in the case of child abuse, it is the victim of the abuse and not the abuser who will have the data coded. New methods designed for EHR data may be able to better control for the peculiar confounding effects of observational data.

There are significant bioethical considerations regarding the use of the RIFTEHR method, including how best to balance the competing demands of protecting patients' privacy with clinicians' duty to warn relatives of potential genetic risks. The method could readily be applied in EHR systems, such that clinicians could easily access the health information of a patient's family members. In the United States, accessing a family member's health information in this manner may be considered a violation of the 1996 Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[21]. On the other hand, case law in the United States has established that healthcare providers have a responsibility to inform a patient's relatives about heritable conditions that may reasonably put the relatives "at risk of harm" [22]. These conflicts may need to be resolved before automatic relationship inference can be used clinically.

We have described and validated a novel method for identifying familial relationships in patient's medical records, and used 4.7 million relationships inferred from the EHRs at two academic medical centers to estimate heritability of disease. We found that heritability estimates were concordant across the two centers, suggesting that the method may have broad applicability. An EHR that is linked to genetic information enables personalized disease risk prediction and facilitates heritability determination for EHR-captured phenotypes that have not been previously studied by family-based or twin studies. Identifying familial relationships is useful for all aspects of medicine, ranging from genetic research to clinical practice, making RIFTEHR a valuable tool for the advancement of precision medicine. The correspondence of our heritability estimates with family based estimates provides a direct and novel validation of the value of electronic health records in making inferences about disease which is now emerging as a central approach in precision medicine.

Materials and Methods

The data for this study was obtained from the inpatient EHR used at the hospitals affiliated with two large academic medical centers in New York City: Columbia University Medical Center and Weill Cornell Medical College. These hospitals operate together as NewYork-Presbyterian Hospital and herein, we will refer to the hospitals and the data associated with them as Columbia and Cornell, respectively.

1. Relationship Inference from the Electronic Health Record (RIFTEHR)

This research was approved by the institutional review boards at the two study sites. As is common practice, when patients received care at either site, they were asked to provide

information about an emergency contact. This information included the person's name, address, phone number, and their relationship to the patient (e.g., parent, sibling, friend). We used the emergency contact information to identify familial relationships in the EHR in cases where the emergency contact person had his or her own record generated by an encounter with the healthcare system. Algorithmically, we then inferred additional relationships from the connectedness of the identified individuals. This information was validated against genetic data and a separate module of the EHR which documented the linkage between mother's and their newborn's medical record. Using the relationships identified, we assigned phenotypes using clinical history, and subsequently evaluated familial recurrence for all available clinical phenotypes.

1.1. Deriving familial relationships from emergency contact data

1.1.1. *Matching emergency contact to medical records.* Our algorithm creates for each patient a list of all reported emergency contacts. Then, for each emergency contact, it attempts to identify a medical record by matching first name, last name, primary phone number and ZIP code. First we consider all cases with first name and filter the table that contains all patients' information to identify records that contain the same first name. We then return the identified records and perform the same comparison with last name, primary phone number and ZIP code. Subsequently, we compare the combination of two variables at a time (i.e. first name and last name, first name and primary phone number, first name and ZIP code, etc.). We then perform combinations of three variables and then of all four variables. We only consider it successful when we identify a single patient that matches to the emergency contact information given. We also capture which variables were used in the matching process for each one of the emergency contacts (i.e. first name and last name; first name, last name and phone number, etc.). The output of this algorithm contains the patient's identifier, the relationship between the patient and the matched emergency contact, the emergency contact's identifier, as well as a list of the variables used to perform the matching process. We use as patient identifiers the Enterprise Master Patient Index (EMPI), when available or the medical record number (MRN). EMPIs are a unique identifier created to refer to multiple MRNs across the healthcare organization. Using EMPIs allow us to perform better in the matching process since duplicates from patients having more than one MRN are excluded.

1.1.2. *Quality Control of matches.* Once the matches are identified, we exclude patients with non-biological relationships (i.e. spouse, friend). Specific relationships are mapped to relationship groups (e.g. the relationship "mother" is mapped to "parent"). We then calculate the age difference between two related patients and exclude parents that are less then 10 years older than their children, children that are less than 10 years younger than their parents, grandparents that are less than 20 years older than their grandchildren, grandchildren that are less than 20 years younger than their grandparent. Since parents and grandparents must be older than their children and grandchildren, we also flip relationships when the age difference between parent or grandparent and its child or grandchild is negative, specifically the relationship "parent" becomes "child" and the relationship "grandparent" becomes "grandchild". The same process is done when the age difference between children and grandchildren in positive. We also exclude every patient that matches to 20 or more distinct emergency contacts. Finally, we generate the opposite relationship for every relationship pair. For example, if we have that A is parent of B, the opposite relationship is that B is child of A.

1.1.3. *Inferring familial relationships.* Using the matches identified, we infer additional relationships. The inference process is made based on familial relationship rules. For example, if patient A is mother of patient B and patient B is mother of patient C, then by inference we

know that A is grandmother of C and C is grandchild of A. The rules used to perform these inferences are described on Supplementary Table 4.

1.1.4. Quality Control of inferred relationships. Once additional relationships are inferred, we remove ambiguous relationships such as "Parent/Aunt/Uncle" if the same pair contains a unique specific relationship, in this case, either "Parent" or "Aunt/Uncle". The same is done for "Child/Nephew/Niece", "Sibling/Cousin", "Parent/Parent-in-law", "Child/Child-in-law", "Grandaunt/Granduncle/Grandaunt-in-law/Granduncle-in-law", "Grandchild/Grandchild-in-law", "Grandnephew/Grandniece/Grandnephew-in-law/Grandniece-in-law", "Grandparent/Grandparent-in-law", "Great-grandchild/Great-grandchild-in-law", "Great-grandparent/Great-grandparent-in-law", "Nephew/Niece/Nephew-in-law/Niece-in-law", and "Sibling/Sibling-in-law".

1.1.5. Identification of families. To identify families in the datasets, we exclude all non-biological relationships such as spouses and in-laws, as well as ambiguous relationships such as "Parent/Parent-in-law". Using both provided and inferred relationships, we created a network where each node corresponds to a patient and edges represent familial relationships. To identify different families, we decomposed network into individual connected components.

1.1.6. Identification of twins. To identify twins we matched siblings that shared the same last name and the same date of birth. We do not have enough information to distinguish between monozygotic and dizygotic twins.

1.2. Evaluation of automatically inferred relationships

1.2.1. Evaluation using the EHR's mother-baby linkage. We used the EHR's mother-baby linkage as gold standard to evaluate identified maternal relationships. We consider true positives cases where maternal relationships present in the EHR's mother-baby linkage table and also identified by our algorithm, false positives when we identified maternal relationships that are discordant with the one in the EHR's mother-baby linkage and lastly, false negatives when a maternal relationship was captured by the EHR's mother-baby linkage but not by our method. Overall performance was evaluated by calculating overall sensitivity and positive predictive value (PPV). In order to assess if matches identified by different variables perform differently, we also computed sensitivity and PPV stratifying the matches by the number of variables used to match the emergency contact to a patient in our healthcare system (Table S2), as well as by the combination of variables (i.e. last name only, first name and last name, etc.) used to perform the match (Table S3).

1.2.2. Evaluation using genetic data with analysis for kinship. Genotype data was collected from existing sources for 186 individuals. Data was collected from three separate sources, the Institute for Genomic Medicine, The Columbia University Medical Center Pathology Department, and the Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project, using whole exome sequencing, Affymetrix CytoScan HD array, and the Illumina Multi-Ethnic Genotyping Array, respectively. In order to select SNPs for kinship, minor allele frequency was filtered to >5%, and genotyping rate to 99% using PLINK [21]. Independent SNPs were selected using the sliding window (100 SNPs) linkage disequilibrium approach. This resulted in a total of 24,752 variants from the Institute for Genomic Medicine data, 8,544 SNPS from the WICER data, and 32,938 SNPs from the Pathology Department data. PLINK was then used to calculate identify by descent by determining $\hat{\pi}$ results (P(IBD=2)+0.5*P(IBD=1)(proportion IBD)) for each pair of individuals. We consider that the predicted relationship is correct if the blood relationship fraction between the two people is the

same as the one expected for the predicted relationship with a margin of error of 20% of the expected blood relationships. For example, for predicted mother-child pairs, two individuals in a pair share 50% (±10%) of their genetic information, then that gives us evidence to consider that the predicted relationship is correct. Likewise, for a predicted aunt-niece pair, the two individuals are expected to share 25% (±5%). Performance was evaluated by calculating PPV.

1.2.3. Evaluation using clinical data. As a qualitative validation of all relationship types, including distant relationships such as great-grandparent, we calculated age difference between all pairs of family relatives and stratified it by relationship type. We compared the identified age differences to what would be expected in a real family structure. For example, great-grandparents should be much older than their great-grandchildren.

2. Phenotyping in the EHR

We used clinical pathology reports as quantitative traits and diagnosis billing codes as dichotomous traits in our study. We extracted the top used clinical pathology reports and mapped them to LOINC codes so that they could be matched between institutions. Each patient may have multiple lab reports over time. To get a single phenotype value we collapsed all reports for each patient into a single value using the mean. This mean represents the average value for the report for the patient over all time available. For example, a patient's mean blood glucose value over their lifetime.

For dichotomous traits we used any diagnosis billing code that was used for at least 1,000 distinct patients. Any patient with evidence of that code in their medical record history was considered a "case." Controls were chosen as any patient that did not have that diagnosis nor any diagnosis that shared an ancestor according to the Clinical Classifications Software (CCS). This tool was developed by the Agency for Healthcare Research and Quality (AHRQ). CCS is composed of diagnoses and procedures organized in two related classification systems. In this study, we only used the diagnoses classifications. The single-level system consists of 285 mutually-exclusive diagnosis categories. It enables researchers to map any of the 3,824 ICD9-CM diagnosis codes into one of the 285 CCS categories. CCS also has a multi-level system composed of 4 levels representing a hierarchy of the 285 categories. The first level is broken into 18 categories. To define a control group, we linked the ICD9 codes associated to a phenotype of interest to their CCS categories using the top-level hierarchical categories. We also generated a table associating each patient to CCS categories they were diagnosed with. Once this mapping was done, each phenotype was associated to one or multiple distinct CCS categories. We matched these CCS categories in the multi-level system to identify the first level parent category. We considered these top level categories as our exclusion criteria: the control cohort for this phenotype should have no mention of any CCS under these categories in its medical records. For example, the controls for atrial fibrillation will exclude patients with cardiovascular diseases.

We semi-manually curated a set of 85 phenotypes to use for training and testing the SOLARStrap algorithm (*See Methods 3.3*). For these 85 phenotypes, we grouped closely related diagnoses codes together to increase the total number of patients (Table S6).

3. Estimation of heritability from the Electronic Health Records

3.1. Rationale

The primary and most significant challenge when using traits defined from an observational resource, like the electronic health records (EHR), is the lack of ascertainment. In a heritability study, the phenotype of each study participate is, ideally, carefully evaluated and quantified.

This is infeasible, however, when the cohort contains millions of patients with thousands of phenotypes. The differential probability that a given individual will be phenotyped for a study trait is the *ascertainment bias*. The bias may depend on many latent factors, including the trait being studied, the trait status of relatives, the proximity to the hospital, and an individual's ethnicity and cultural identification, among others. The consequence of this uncontrolled ascertainment bias is that heritability estimates will be highly dependent on the particular individuals in the study cohort. We used repeated sub-sampling to characterize this dependency quantitatively. We define the observational heritability, or h_2^o, as the average of the statistically significant sample estimates (using median). For a given trait, the procedure, which we call SOLARStrap, involves sampling families, running SOLAR to estimate sample heritability, and rejecting or accepting the estimate based on a set of quality control criteria. Each step is detailed below.

3.2. SOLARStrap Protocol

3.2.1. Building pedigree files. Of the 223,307 families at Columbia there were 6,894 that contained conflicting relationships -- where two individuals were inferred to have two different relationships. At Cornell 3,258 families of 155,811 contained conflicts. These families were excluding from the heritability studies. In some cases, more than one mother or father is annotated for an individual. This could be because of duplicate patient records or errors in the EHR relationship extraction. We resolve these issues by choosing the mother or father that has more relationships in the family. The other relationship is discarded. We then constructed a master pedigree file for each site. To construct this pedigree file we iterate through each member of each family. For each individual we will either know the mother and father from the EHR derived relationships or not. If not known, then a new identifier is created to represent the parent. At this point we iterate through all other family members and record the relationships between the new individual and each family member. We repeat this process until the entire pedigree file is created. The master pedigree files contain 1,377,173 and 940,040 individuals for Columbia and Cornell, respectively.

3.2.1. Sampling Families. The number of families that are sampled combined with the prevalence of the trait defines the power of the heritability analysis. A smaller heritability can be detected with larger sample sizes. However, as the sample size increases toward the total number of families the variance in heritability that can be observed will decrease. This is because we are sampling without replacement. Since we do not know a priori what the magnitude of the heritability will be or what the variance will be we iterate through sample sizes from 100 to the total number of available families. The maximum sample size is defined by the limitations of SOLAR which can only handle a maximum of 32,000 individuals per pedigree file. For each sample size we perform 200 samplings. For each of these we build a custom pedigree and phenotype files and run SOLAR to estimate the heritability. We then aggregate the results.

3.2.2. Sample pedigree files. For each sampling a set of N families are selected. To construct the sample pedigree file we identify all lines from the master pedigree files that correspond to these families and create a new file from this subset.

3.2.3. Sample phenotype files. Once the pedigree file is created, we iterate over every individual in the pedigree and use the reference trait data and demographic data to enter the phenotype status and age of the patient. If no phenotype data are available for the individual we enter it as missing. For dichotomous traits the trait values are either 0 (absence), 1 (presence), or missing and a "proband" is randomly assigned by selected a single individual from each family that has the trait. See "Phenotyping in the EHR" for a description of how these traits are assigned. For quantitative traits we enter the quantitative value or missing.

3.2.2. Running SOLAR. We use SOLAR to estimate both quantitative and dichotomous trait heritability using a mixed linear model. In both cases sex and age are modeled as covariates. After the pedigree and phenotype files are loaded the heritability is estimated with the `polygenic -screen` command. We used the tdist command in SOLAR to adjust quantitative traits that are not normally distributed. For dichotomous traits one "proband" is chosen at random for each family. SOLAR will automatically detect the presence of a dichotomous trait and convert the estimate from the observed scale to the liability scale. The heritability, error on the heritability, and the p value are saved from each run for later analysis and aggregation.

3.2.3. Quality Control of SOLAR heritability solutions. SOLAR does not converge on a solution for heritability for all samples. Errors in the pedigree or in the ascertainment of phenotypes are the most likely causes for these failures. First, we reject any runs of SOLAR that result in no solution for the heritability. We then consider two additional criteria that must be met in order for a solution to be considered legitimate: (i) *edge epsilon* (ϵ_e), any estimate within ϵ_e of 1 or 0 is rejected; and (ii) *noise epsilon* (ϵ_n), any estimate with implausibly low error is rejected (h_2 error is less than ϵ_n of the h_2 estimate). These hyperparameters are set using a set of phenotypes for which we have observational heritability estimates and high confidence literature reported heritabilities from other studies.

POSA. After filtering the SOLAR solutions for the basic criteria, we define an additional quality control metric called the Proportion Of Significant Attempts, or POSA. POSA is defined as the number of solutions with a p value less than $\boldsymbol{\alpha_{POSA}}$ divided by the total number of converged solutions (or attempts). The POSA is important because it is closely related to the power of the analysis. A fully powered analysis will have a POSA of 1, meaning that all of the converged estimates are statistically significant. A POSA of 0.5 means that only half of the converged estimates are statistically significant. When the families were sampled the observed heritability was large enough to be detected with $p < \boldsymbol{\alpha_{POSA}}$ half of the time. In other words, we were powered to detect a heritability in 50% of samplings. We show that the higher the POSA, the more accurate the heritability estimates are (Figure S2). We chose a minimum POSA score, $POSA_{lower}$ and the $\boldsymbol{\alpha_{POSA}}$ using a set of phenotypes for which we have observational heritability estimates and high confidence literature reported heritabilities.

3.2.4. Aggregation of sampling results (computing h_2^o). For each sampling that passes quality control and meets the minimum POSA score, we compute the h_2^o as the median. The median h_2^o corresponds to a single run of SOLAR that has passed all of the quality control filters. We used the standard error reported by SOLAR for that run as the error of the h_2^o. We found that this error is closely related to the sampling variance (Figure S3). All raw heritability estimates that pass the initial quality control are made publicly available for reanalysis.

3.3. Fit and validation of hyperparameters

Heritability estimates for 91 phenotypes were mined from the literature along with their corresponding confidence intervals, if they were available. We performed a brute force search through the parameter space. Possible values for edge epsilon (ϵ_e) were (0, 1e-9, 1e-8, and 1e-7). Possible values for noise epsilon (ϵ_n) were (0.01, 0.025, 0.05, 0.075, 0.1, and 0.2). Possible values for $\boldsymbol{\alpha_{POSA}}$ were (0.03, 0.05, 0.1, 0.25, 0.5, and 1.0). Possible values for the $POSA_{lower}$ were (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.975). We evaluated each set of parameters for the correlation between the h_2^o and the h_2. Only a single site was used to fit these parameters, leaving data from the other site available for validation. The maximum correlation was 0.558 with $\boldsymbol{\alpha_{POSA}}$ = 0.05, ϵ_e = 1e-9, ϵ_n = 0.05, and $POSA_{lower}$ = 0.7. At these parameter settings 19 traits passed quality control. The average difference between h_2^o and h_2 was 17.7% ± 8%.

To evaluate the generalizability of the hyperparameters, we applied them to the validation site data. 10 traits passed the quality controls and we found that they were correlated with literature estimates of heritability (r = 0.73, p = 0.016). The average difference between h_2^o and h_2 was 7.6% ± 9%.

3.4. Preparation of data for analysis on external computing clusters

Due to the high number of heritability estimates that need to be computed, external computing resources are used: The Open Science Grid (OSG) and Amazon Web Services (AWS). The Open Science Grid (OSG) is a massive computing resource funded by the Department of Energy and the National Science Foundation. The OSG is comprised of over 100 individual sites throughout the United States, primarily located at universities and national laboratories. The sites contain anywhere from hundreds to tens of thousands of CPU cores available for scientific research[23,24]. AWS is used to supplement this resource, which makes available on-demand compute instances with high performance capacity. Per institutional requirements, no protected health information or personally identifying information can be transferred to systems outside of our institutional networks. To leverage these resources for our computing task we prepared a data subset according to the Safe Harbor guidance provided by the U.S. Department of Health and Human Services (http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html). Here is a point-by-point account of how we processed the data for Safe Harbor for each of the 18 identifiers: (A) we removed first, middle, and last names for all patients, (B) all patient address information is removed, (C) all dates are removed and all ages over 89 are coded as "90", (D) telephone numbers and (E) fax numbers are removed, (F) there are no email addresses in our subset of the clinical data, (G) there are no social security numbers in our subset of the clinical data, (H) medical record numbers are mapped to a 10 digit random number and the mapping is stored on a limited access PHI-certified server within the institutional firewall and will never be made available, (I) there are no health plan beneficiary numbers in our data subset, (J) there are no account numbers in our data subset, (K) there are no certificate or license numbers, (L) there are no vehicle numbers or serial numbers in our data subset, (M) there are device identifiers or serial numbers, (N) there are no URLs in our data subset, (O) there are no IP addresses in our data subset, (P) there are no biometric identifiers in our data subset, (Q) there are no full-face or comparable images in our data subset, (R) there are no other uniquely identifying characteristics or numbers. All data were transferred using secure file transfer protocols using encryption and were destroyed immediately after retrieval of the results. In total we used over 20,000 cpu-hours to compute heritability estimates for 2,089 traits.

3.5. Investigation of Heritability of "Victim of Child Abuse"

The trait with the highest heritability in our study was "victim of child abuse" coded as V61.21. The heritability was 0.90 with the 95% confidence interval spanning from 0.73 to 1.00 at Columbia when sampling 600 families. There was not enough data at Cornell to estimate the heritability. At Columbia this trait was coded for 946 families where at Cornell it was available for only 134 families. None of the estimates from Cornell passed our primary QC screen. At Columbia, however, estimates are available and sampling sizes of 200, 300, 400, 500, and 600 all passed the second QC stage (POSA > 0.7). The heritability estimates ranged from 0.76 (0.45-0.97) to 0.90 (0.73-1.000) and are shown in Table S5. This trait is not actually a trait of the individual that the code is assigned, but to another individual with whom the patient interacts. We suspected that the high heritability may be an artifact of multiple siblings in a single family being abused by a single individual. To account for this, we chose only a single affected sibling for each family. All other siblings were coded as having their trait "missing." We then ran SOLARStrap for sample sizes of 200, 300, 400, 500, and 600 for comparison (Table S5).

4. Estimation of sibling and familial recurrence for dichotomous traits

We estimated sibling recurrence as the proportion of individuals with that trait given that they have a sibling with the trait. We randomized the choice of primary sibling that the probability is conditioned upon. We computed familial recurrence similarly, except that any relationship type was allowed. Both sibling and familial recurrence were only calculated for conditions with 10 or more concordant pairs. The recurrence rate is calculated by $\frac{2*Concordant\ pairs}{2*Concordant\ pairs+Discordant\ pairs}$ and the error by $\sqrt{\frac{recurrence*(1-recurrence)}{2*concordant\ pairs+discordant\ pairs}}$.

To compare disease recurrence rates, we computed recurrence for each relationship type. To test if the groups were statistically different, we performed a Chi-squared test with Bonferroni correction.

5. Preparation of clinical data for release

Due to institutional restrictions, we cannot release the exact data as it was used in our analysis. However, we are sensitive to issues regarding reproducibility and replicability. Therefore, we have modified the dataset according to the rules of Safe Harbor as provided by the U.S. Department of Health and Human Services. The processing of the data for release was performed as described in section 3.4. However, in this case we took three additional precautions beyond what is required for Safe Harbor since these data will be made completely public. We are releasing data for a single trait (rhinitis). We will continue to release more traits as the data are reviewed to protect patient privacy with the ultimate goal of releasing all of the trait and relationship data for all phenotypes. No data are released for families containing more than five members. This will protect against identification through unique familial relationships situations. All aggregate data and their corresponding statistics are released without obfuscation. The data are available on the supporting website: http://riftehr.tatonettilab.org/.

6. Computational and statistical software

Statistical analysis, data preparation, and figure creation was performed using Python 2.7. The python system environment is described fully in the supplemental materials. Relationship inferences was implemented in Julia 0.4.3. All correlations are reported as Pearson correlation coefficients, unless otherwise noted. All code for RIFTEHR and SOLARStrap is available on the supporting website: http://riftehr.tatonettilab.org/.

7. Literature review

For validation purposes, we performed literature review on heritability estimates on 128 traits. We started by analyzing studies that were included in the table available at http://www.snpedia.com/index.php/Heritability (accessed on March 2016). We then downloaded all papers we had access to and extracted the described trait with the respective heritability estimates as well as the confidence intervals, when available.

Acknowledgements

FP and DKV are supported by R01H5021816. KQ, RV, AY, and NPT are supported by R01GM107145. KK is supported by NIDDK R01DK105124. RV, KK, and NPT are supported by the Herbert Irving Scholars Award. GH is supported by R01LM006910. This research used resources from the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. Collection of genetic samples was supported by R01HS022961.

References:

1. Guttmacher AE, Collins FS, Carmona RH. The family history--more important than ever. *N Engl J Med*. 2004;351(22):2333-2336. doi:10.1056/NEJMsb042979.

2. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics*. 2013;14(2):139-149. doi:10.1038/nrg3377.

3. Mucci LA, Hjelmborg JB, Harris JR, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA*. 2016;315(1):68-76. doi:10.1001/jama.2015.17703.

4. Polderman TJC, Benyamin B, de Leeuw CA, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Publishing Group*. 2015;47(7):702-709. doi:10.1038/ng.3285.

5. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*. 2013;20(1):144-151. doi:10.1136/amiajnl-2011-000681.

6. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association : JAMIA*. 2013;20(1):117-121. doi:10.1136/amiajnl-2012-001145.

7. Mayer J, Kitchner T, Ye Z, et al. Use of an electronic medical record to create the marshfield clinic twin/multiple birth cohort. *Genet Epidemiol*. 2014;38(8):692-698. doi:10.1002/gepi.21855.

8. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*. 1998;62(5):1198-1211. doi:10.1086/301844.

9. Pietiläinen KH, Söderlund S, Rissanen A, et al. HDL subspecies in young adult twins: heritability and impact of overweight. *Obesity (Silver Spring)*. 2009;17(6):1208-1214. doi:10.1038/oby.2008.675.

10. Souren NY, Paulussen ADC, Loos RJF, et al. Anthropometry, carbohydrate and lipid metabolism in the East Flanders Prospective Twin Survey: heritabilities. *Diabetologia*. 2007;50(10):2107-2116. doi:10.1007/s00125-007-0784-z.

11. Smearman EL, Almli LM, Conneely KN, et al. Oxytocin Receptor Genetic and Epigenetic Variations: Association With Child Abuse and Adult Psychiatric Symptoms. *Child Dev*. 2016;87(1):122-134. doi:10.1111/cdev.12493.

12. Palmier-Claus J, Berry K, Darrell-Berry H, et al. Childhood adversity and social functioning in psychosis: Exploring clinical and cognitive mediators. *Psychiatry Res*. 2016;238:25-32. doi:10.1016/j.psychres.2016.02.004.

13. Goode WJ. Force and Violence in the Family. *Journal of Marriage and the Family*. 1971;33(4):624. doi:10.2307/349435.

14. van Beijsterveldt CEM, Boomsma DI. Genetics of parentally reported asthma, eczema and rhinitis in 5-yr-old twins. *European Respiratory Journal*. 2007;29(3):516-521.

doi:10.1183/09031936.00065706.

15. Lorberbaum T, Sampson KJ, Woosley RL, Kass RS, Tatonetti NP. An Integrative Data Science Pipeline to Identify Novel Drug Interactions that Prolong the QT Interval. *Drug Saf*. 2016;39(5):433-441. doi:10.1007/s40264-016-0393-1.

16. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. *Sci Transl Med*. 2012;4(125):125ra31-125ra31. doi:10.1126/scitranslmed.3003377.

17. Boland MR, Shahn Z, Madigan D, Hripcsak G, Tatonetti NP. Birth month affects lifetime disease risk: a phenome-wide method. *Journal of the American Medical Informatics Association : JAMIA*. 2015;22(5):1042-1053. doi:10.1093/jamia/ocv046.

18. Ritchie MD, de Andrade M, Kuivaniemi H. The foundation of precision medicine: integration of electronic health records with genomics through basic, clinical, and translational research. *Front Genet*. 2015;6:104. doi:10.3389/fgene.2015.00104.

19. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health*. 2015;36:345-359. doi:10.1146/annurev-publhealth-031914-122747.

20. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med*. 2015;7(1):41. doi:10.1186/s13073-015-0166-y.

21. United States. *Health Insurance Portability and Accountability Act of 1996. Public Law 104-191.* Vol 110. 1996:1936-2103.

22. Suarez R. Breaching Doctor-Patient Confidentiality: Confusion among Physicians about Involuntary Disclosure of Genetic Information. *S Cal Interdisc LJ*. 2011.

23. Pordes R, Petravick D, Kramer B, et al. The open science grid. *J Phys: Conf Ser*. 2007;78(1):012057. doi:10.1088/1742-6596/78/1/012057.

24. Sfiligoi I, Bradley DC, Holzman B, Mhashilkar P, Padhi S, Würthwein F. *The Pilot Way to Grid Resources Using glideinWMS*. Vol 2. IEEE; 2009:428-432. doi:10.1109/CSIE.2009.950.

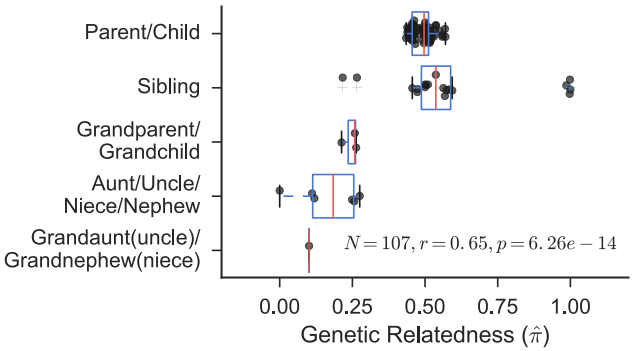
A EHR-derived relationships compared to 79,786 known mother-child relationships

	True Positive	False Positive	False Negative	Sensitivity	PPV
CUMC	35775	1600	2720	0.9293	0.9572
Cornell	37797	657	1237	0.9683	0.9829

B EHR-derived relationships versus genetically derived relationships for 185 individuals and 122 relationships

Relationship	N	Expected	Genetic Relatedness Mean +/- SD	PPV
Parent/Child	78	0.5 +/- 0.1	0.49 +/- 0.03	1.000
Sibling	19	0.5 +/- 0.1	0.59 +/- 0.23	0.684
Excluding Twins	15	0.5 +/- 0.1	0.49 +/- 0.11	0.867
Grandparent/Grandchild	3	0.25 +/- 0.05	0.24 +/- 0.02	1.000
Aunt/Uncle/Niece/Nephew	6	0.25 +/- 0.05	0.17 +/- 0.10	0.500
Grandaunt/Granduncle/ Grandnephew/Grandniece	1	0.06 +/- 0.01	0.10 +/- 0.00	0.000

C



D

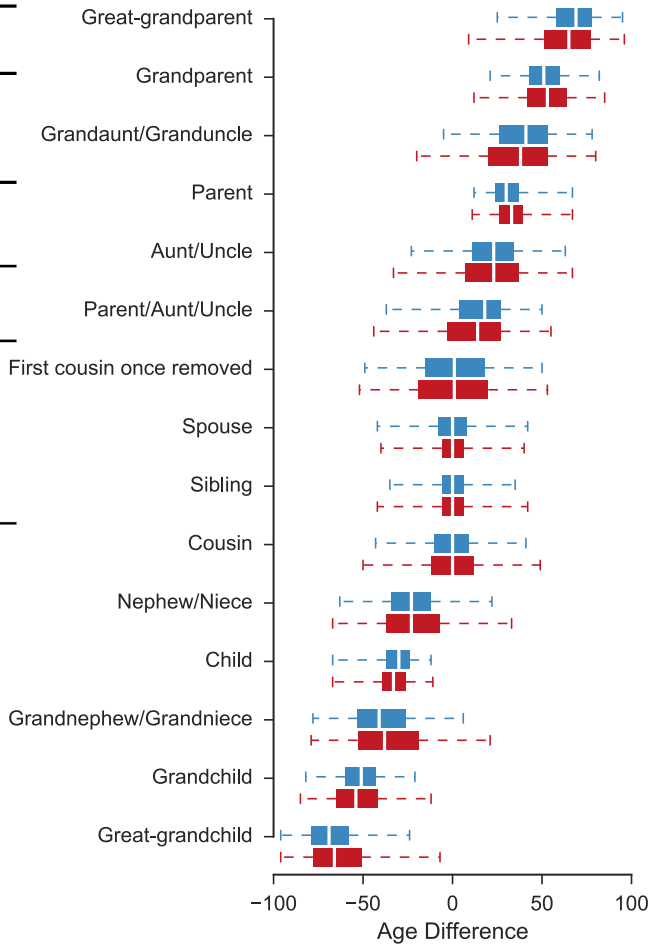


Figure 1. Validation of familial relationships inferred from the EHR. (A) The medical centers at both Columbia and Cornell have implemented a link between the electronic health records of mother and baby at the time of birth. We used these links as a gold standard to evaluate RIFTEHR, our algorithm for automatically inferring relationships from the EHR. Of 40,095 mother-baby links at Columbia, RIFTEHR correctly identifies 35,775, falsely identifies 1,600 and misses 2,720. Positive predictive value (PPV) is 96% and sensitivity is 93%. Of 39,691 mother-baby links at Cornell, RIFTEHR correctly identifies 37,797, falsely identifies 657, and misses 1,237. PPV is 98% and sensitivity is 97%. (B and C) Through biobanks at Columbia, 185 of the patients with identified relationships from RIFTEHR also had genetic data available and appropriately consented for use in our study. For these 185 patients, RIFTEHR predicted a total of 122 relationships: 78 parent/child relationships, 19 sibling relationships, 3 grandparent/grandchild relationships, 6 aunt/uncle/niece/nephew relationships, and one grandaunt/grandniece relationship. Genetic relatedness was determined for each pair of individuals. All 78 parent/child relationships had the expected genetic relatedness of 50% (49%±3%). Of the siblings predicted by RIFTEHR 13 were full siblings, 2 were half siblings (genetic relatedness of 25%), and 4 were identical twins. The high rate of twins in our small sample is a result of the secondary use of existing data – which was originally collected for genetic studies. Excluding these twins yields a more accurate estimate of RIFTEHR's performance (PPV=87%). Overall the RIFTEHR relationship and the genetic relationship were significantly correlated (r=0.65, p=6.24e-14). (D) Average age differences for each relationship type. We computed the age differences for each pair of individuals at both Columbia (blue) and Cornell (red). The age differences are consistent across sites.

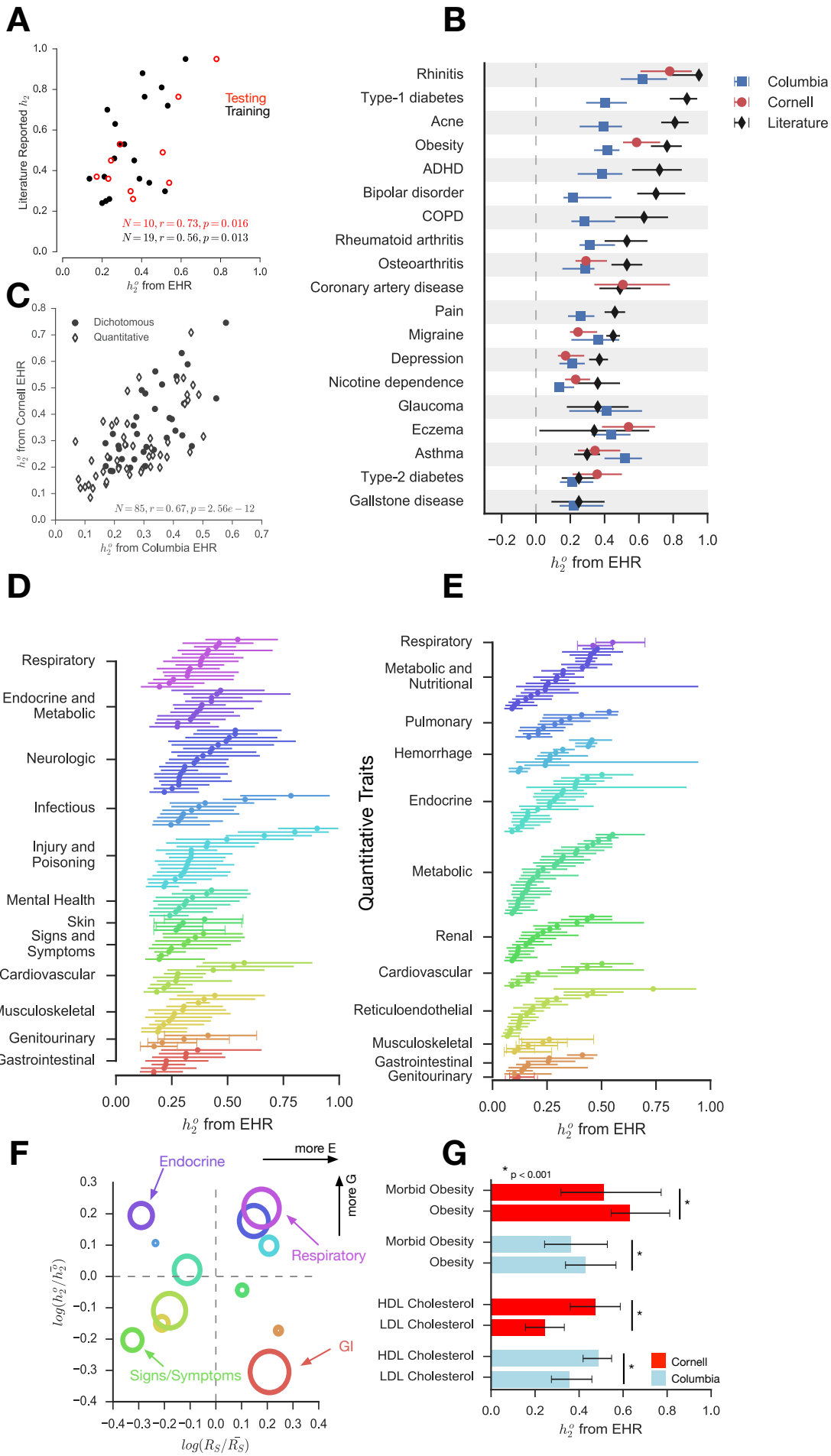


Figure 2. Estimating heritability of disease using electronic health records. We designed a method, called SOLARStrap, for estimating the heritability of traits where the phenotype is derived under unknown ascertainment biases, the h_2^o. We trained the hyperparameters of the model on a small subset of manually defined phenotypes with available heritability estimates from the literature (*Materials and Methods*) using data from Columbia and tested these parameters at Cornell. (A) We found that performance was consistent across both sites and that h_2^o is significantly correlated with literature estimates of h_2. (B) Comparison of h_2^o (from Columbia and Cornell EHR) and h_2 (from Literature) for 28 traits used for fitting the hyperparameters. Median and 95% confidence interval are shown. (C) We evaluated the heritability of just over 1,494 traits at Columbia and 1,145 traits at Cornell (*Materials and Methods*). We performed the analysis independently each site. After quality control filters we found 216 traits with significant heritability at Columbia and 160 traits at Cornell, with 85 traits falling in the intersection. These 85 traits were significantly correlated between the two sites (r=0.67, p=2.56e-12). (D) 124 dichotomous traits (from disease billing codes) grouped by disease category and sorted by heritability within each group. Disease categories are sorted by the median heritability of the diseases within the category. Respiratory disease has the highest average heritability and gastrointestinal disease has the lowest average heritability. (E) 92 quantitative traits (from clinical pathology reports) grouped by disease category and sorted by heritability within each group. Trait categories are sorted by the median heritability of the traits within the category. Respiratory disorders have the highest average heritability followed by metabolic and nutritional disorders. Gastrointestinal and genitourinary disorders have the lowest average heritability of their corresponding quantitative traits. (F) For the 124 dichotomous traits we have both estimates of heritability and sibling recurrence rates. The median heritability and recurrence rate were computed for each category and then normalized to the overall median heritability across all groups (y axis). The same was done for recurrence rates (x axis). Each category is shown as an open circle colored according to (D). The size of the circle indicates the number of traits within that category. Categories in the top right quadrant have higher than average heritability and higher than average recurrence rates while categories in the top left quadrant have low recurrence rates and high heritability, etc. (G) Observational heritability for morbid obesity and obesity at Columbia (light blue) and Cornell (red) as well as for for HDL cholesterol at Columbia (light blue) and Cornell (red).

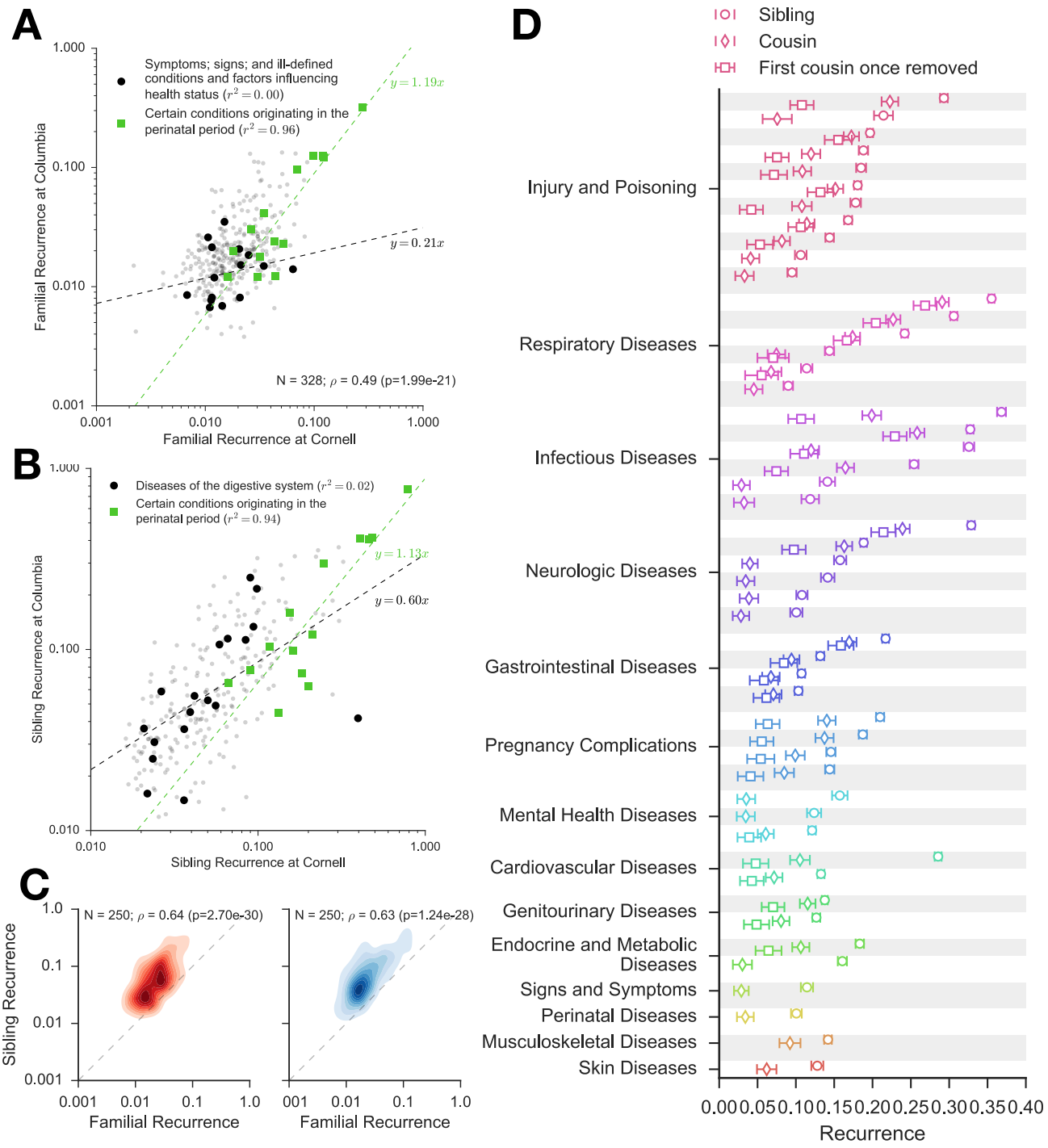


Figure 3: Estimating Familial and Sibling Recurrence Rates using EHR data. (A) Correlation of familial recurrence estimates for 328 conditions (rho = 0.49, p = 1.99e-21) between Cornell and Columbia. Perinatal conditions (green) represent the most concordant (r^2 = 0.96) and signs and symptoms (black) represent the least concordant (r^2 = 0). (B) Correlation of sibling recurrence estimates for 250 conditions between Cornell and Columbia (rho = 0.71, p = 2.52e-40). Perinatal conditions (green) once again represents the most concordant (r^2 = 0.94) and diseases of the digestive system (black) represent the least concordant (r^2 = 0.02). (C) Sibling recurrence estimates versus familial recurrence estimates at Columbia (left, blue) and Cornell (right, red). Sibling recurrence and familial recurrence is significantly correlated at both sites and, on average, is greater than familial recurrence at both sites. (D) Sibling recurrence by disease category stratified by relationship type (sibling, cousin, first cousin once removed) for Columbia. Circles represent sibling recurrence rates, diamonds represent cousin recurrence rates and squares represent first cousin once removed along with the 95% confidence interval.

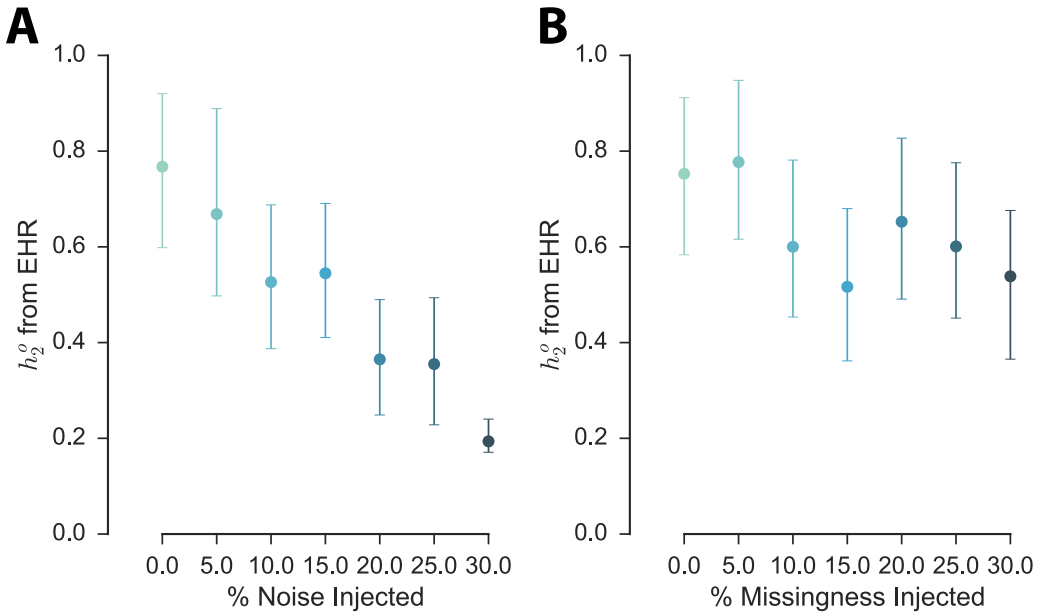
Variable	Columbia	Cornell
N	682,267	437,375
Relationships	3,244,380	1,534,760
N provided relationships	488,932	297,011
N inferred relationships	2,755,448	1,237,749
Gender, Female	418,657 (61.29%)	261,482 (59.74%)
Age	40.15 (24.81)	39.85 (25.02)
Race/Ethnicity		
Black or African American	69,506 (10.19%)	30,975 (7.08%)
White	123,800 (18.15%)	110,485 (25.26%)
Hispanic or Latino	373,552 (54.75%)	52,087 (11.91%)
Other	11,438 (1.68%)	26,687 (6.10%)
Unknown/Declined to answer	103,971 (15.24%)	217,141 (49.65%)
Degree of relationship		
First (i.e. child, parent, sibling)	1,388,858	814,650
Second (e.g. grandchild)	605,922	225,796
Third (e.g. great-grandparent)	432,262	137,712
Fourth (e.g. great-great-grandchild)	215,300	61,986
Other		
None (e.g. spouse, in-laws)	172,158	127,748
Unknown (e.g. parent/parent-in-law)	429,880	166,868

Table 1. Demographic data of the electronic health records at the medical centers of Columbia and Cornell University.

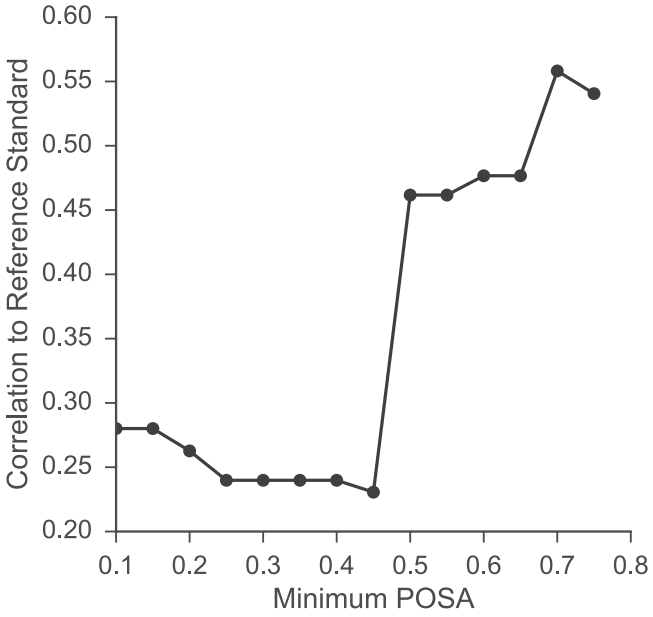
Dichotomous Disease Category	Median h_2^o (min-max)	Trait with Highest Heritability			Trait with Lowest Heritability		
		ICD9 Code	Name	Median h_2^o (95% CI)	ICD9 Code	Name	Median h_2^o (95% CI)
Respiratory Diseases	0.38 (0.19-0.55)	477.9	Allergic rhinitis	0.55 (0.40-0.73)	786.2	Cough	0.19 (0.11-0.35)
Endocrine and Metabolic Diseases	0.38 (0.27-0.47)	250.01	Insulin dependent diabetes mellitus	0.47 (0.25-0.67)	268.9	Vitamin D deficiency	0.27 (0.15-0.46)
Neurologic Diseases	0.36 (0.22-0.53)	367.1	Myopia	0.53 (0.36-0.74)	780.4	Giddiness	0.22 (0.15-0.37)
Infectious Diseases	0.34 (0.25-0.78)	79.89	Specific viral infections	0.78 (0.56-0.96)	41.86	Helicobacter-associated disease	0.25 (0.16-0.42)
Injury and Poisoning	0.34 (0.22-0.90)	V61.21	Victim of child abuse	0.90 (0.73-1.00)	V82.9	Screening for disorder	0.22 (0.13-0.28)
Mental Health Diseases	0.31 (0.24-0.43)	309.28	Adjustment disorder with mixed emotional features	0.43 (0.30-0.59)	300	Anxiety	0.24 (0.15-0.33)
Skin Diseases	0.29 (0.27-0.40)	706.1	Acne	0.40 (0.22-0.57)	682.9	Abscess	0.27 (0.17-0.49)
Signs and Symptoms	0.28 (0.19-0.39)	919.4	Nonvenomous insect bite	0.39 (0.21-0.57)	883	Open wound of finger without complication	0.19 (0.13-0.40)
Cardiovascular Diseases	0.27 (0.18-0.57)	411.1	Preinfarction syndrome	0.57 (0.39-0.88)	786.59	Chest pain	0.18 (0.12-0.35)
Musculoskeletal Diseases	0.26 (0.19-0.44)	727.3	Inflammation of bursa	0.44 (0.27-0.67)	724.2	Low back pain	0.19 (0.11-0.32)
Genitourinary Diseases	0.26 (0.17-0.41)	611.72	Breast irregular nodularity	0.41 (0.22-0.63)	599	Urinary tract infectious disease	0.17 (0.11-0.28)
Gastrointestinal Diseases	0.22 (0.17-0.37)	533.9	Peptic ulcer without hemorrhage, without perforation AND without obstruction	0.37 (0.20-0.65)	535	Acute gastritis	0.17 (0.12-0.31)

Quantitative Disease Category	Median h_2^o (min-max)	Trait with Highest Heritability			Trait with Lowest Heritability		
		LOINC Code	Name	Median h_2^o (95% CI)	LOINC Code	Name	Median h_2^o (95% CI)
Respiratory Disorders	0.51 (0.46-0.55)	19213-8	pH of Mixed venous blood	0.55 (0.47-0.70)	11558-4	pH of Blood	0.46 (0.39-0.55)
Metabolic and Nutritional Disorders	0.29 (0.09-0.48)	2573-4	Lipoprotein.alpha [Mass/volume] in Serum or Plasma	0.48 (0.41-0.56)	5810-7	Specific gravity of Urine by Refractometry	0.09 (0.05-0.14)
Pulmonary Disorders	0.29 (0.17-0.54)	19223-7	Carbon dioxide, total [Moles/volume] in Mixed venous blood	0.54 (0.47-0.58)	14627-4	Bicarbonate [Moles/volume] in Venous blood	0.17 (0.10-0.27)
Hemorrhage	0.26 (0.12-0.46)	28542-9	Platelet mean volume [Entitic volume] in Blood	0.46 (0.35-0.55)	20570-8	Hematocrit [Volume Fraction] of Blood	0.12 (0.07-0.16)
Endocrine Disorders	0.26 (0.09-0.50)	19123-9	Magnesium [Mass/volume] in Serum or Plasma	0.50 (0.40-0.65)	5810-7	Specific gravity of Urine by Refractometry	0.09 (0.05-0.14)
Metabolic Disorders	0.21 (0.09-0.55)	19213-8	pH of Mixed venous blood	0.55 (0.47-0.70)	5810-7	Specific gravity of Urine by Refractometry	0.09 (0.05-0.14)
Renal Disorders	0.19 (0.09-0.46)	28542-9	Platelet mean volume [Entitic volume] in Blood	0.46 (0.35-0.55)	5810-7	Specific gravity of Urine by Refractometry	0.09 (0.05-0.14)
Cardiovascular Disorders	0.19 (0.09-0.50)	19123-9	Magnesium [Mass/volume] in Serum or Plasma	0.50 (0.40-0.65)	5810-7	Specific gravity of Urine by Refractometry	0.09 (0.05-0.14)
Reticuloendothelial Disorders	0.16 (0.07-0.74)	2170-9	Deprecated Cobalamin [Mass/volume] in Serum	0.74 (0.28-0.93)	19048-8	Nucleated erythrocytes/100 leukocytes [Ratio] in Blood	0.07 (0.04-0.12)
Musculo_skeletal System	0.16 (0.10-0.26)	3022-1	Deprecated Thyroxine free index in Serum or Plasma	0.26 (0.12-0.46)	6768-6	Alkaline phosphatase [Enzymatic activity/volume] in Serum or Plasma	0.10 (0.05-0.27)
Gastrointestinal Disorders	0.16 (0.10-0.41)	1751-7	Albumin [Mass/volume] in Serum or Plasma	0.41 (0.34-0.48)	6768-6	Alkaline phosphatase [Enzymatic activity/volume] in Serum or Plasma	0.10 (0.05-0.27)
Genitourinary Disorders	0.11 (0.10-0.12)	2756-5	pH of Urine	0.12 (0.08-0.19)	2160-0	Creatinine [Mass/volume] in Serum or Plasma	0.10 (0.06-0.21)

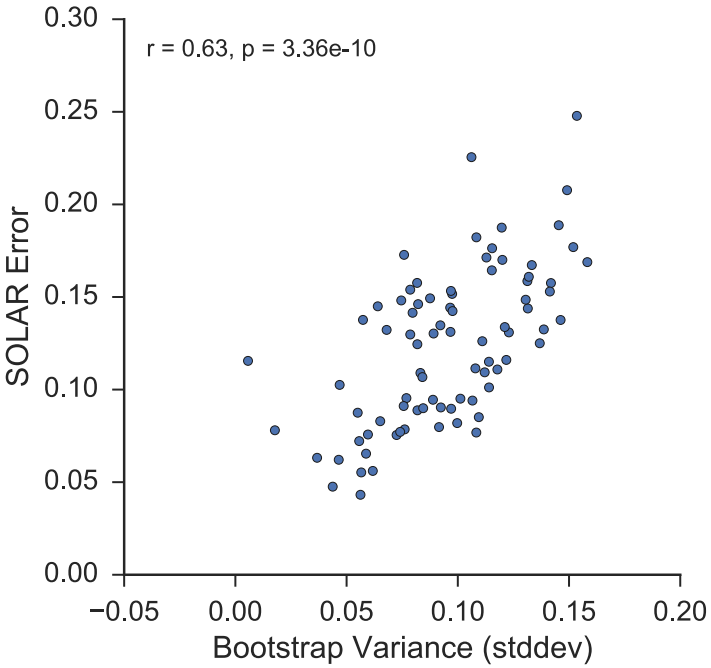
Table 2. Heritability Ranges for Dichotomous and Quantitative Trait Categories. The median observational heritability and ranges are shown for each of the 12 dichotomous trait categories and the 12 quantitative trait categories. Within each category the trait with the highest heritability and the trait with the lowest heritability are shown.



Supplemental Figure 1. SOLARStrap sensitivity analysis. (A) The effect of noise injection on the estimate of observational heritability of rhinitis. We injected noise into the data by randomly shuffling a subset of the patient diagnoses. This simulates misclassification (misdiagnosis or missed diagnosis) in the medical records. When no noise is injected the estimate is 0.77 (0.60-0.92). As noise is introduced the estimate of the heritability decreases to 0.36 (0.23-0.49) once one quarter of the data are randomized. (B) The effect of missingness injection on the estimate of observational heritability of rhinitis. We injecting missingness into the data by randomly removing a subset of the patient diagnoses. This simulates data that are missed by the medical records – an event that is common, especially at tertiary medical centers. When no data are removed the observational heritability is 0.75 (0.58-0.91). The heritability estimate remains consistent until 30% of the data are removed at which time the estimate is 0.54 (0.37-0.68).



Supplemental Figure 2. Accuracy of heritability estimates relies on the proportion of significant attempts (POSA). The POSA score is a measure of how reliable the heritability estimate is that is generated by SOLARStrap. If none of the sample estimates are statistically significant then the POSA will be 0 indicating that the analysis is underpowered. As the sample size increases the power will increase and so does the the POSA score. At a POSA of 0.5 or above, the correlation of the observational heritability estimates to the reference standard jumps significantly. A POSA of 0.7 or above was found to yield the maximum correlation between SOLARStrap heritability estimates and the reference standard.



Supplemental Figure 3. SOLAR error versus SOLARStrap variance. The error estimate from SOLAR is significantly correlated to the sampling variance of the heritability estimates (r=0.63, p=3.3e-10).

Supplemental Table 1. Relationships by degree.

Degree of relationship	Relationship	N Columbia	N Cornell
First	Child	482,308	298,136
	Parent	482,308	298,136
	Sibling	424,242	218,378
Second	Aunt/Uncle	185,822	65,410
	Grandchild	117,139	47,488
	Grandparent	117,139	47,488
	Nephew/Niece	185,822	65,410
Third	Cousin	148,806	37,370
	Grandaunt/Granduncle	96,675	31,764
	Grandnephew/Grandniece	96,675	31,764
	Great-grandchild	45,053	18,407
	Great-grandparent	45,053	18,407
Fourth	First cousin once removed	94,404	19,596
	Great-great-grandchild	17,854	7,531
	Great-great-grandparent	17,854	7,531
	Great-grandaunt/Great-granduncle	42,594	13,664
	Great-grandnephew/Great-grandniece	42,594	13,664
Other			
None	Child-in-law	0	278
	Parent-in-law	0	278
	Spouse	172,158	127,192
Unknown	Aunt/Uncle/Aunt-in-law/Uncle-in-law	13,220	5,234
	Child/Child-in-law	52,186	24,733
	Child/Nephew/Niece	31,818	8,078
	Grandaunt/Granduncle/Grandaunt-in-law/Granduncle-in-law	12,035	4,278
	Grandchild/Grandchild-in-law	12,876	4,578
	Grandnephew/Grandniece/ Grandnephew-in-law/Grandniece-in-law	12,035	4,278
	Grandparent/Grandparent-in-law	12,876	4,578
	Great-grandchild/Great-grandchild-in-law	5,799	2,346
	Great-grandparent/Great-grandparent-in-law	5,799	2,346
	Nephew/Niece/Nephew-in-law/Niece-in-law	13,220	5,234
	Parent/Aunt/Uncle	31,818	8,078
	Parent/Parent-in-law	52,186	24,733
	Sibling/Cousin	41,270	9,142
	Sibling/Sibling-in-law	132,742	59,232

Supplemental Table 2. Performance by number of paths.

	Columbia			Cornell		
N of Paths	**True Positive**	**False Positive**	**PPV**	**True Positive**	**False Positive**	**PPV**
1	4340	1021	0.8096	2979	391	0.884
2	3911	355	0.9168	4114	95	0.9774
3	2438	55	0.9779	4753	53	0.989
4	2696	89	0.968	2089	63	0.9707
5	3075	16	0.9948	4219	29	0.9932
6	5840	30	0.9949	10170	19	0.9981
7	3892	10	0.9974	4100	12	0.9971
8	3105	13	0.9958	1739	19	0.9892
9	2575	6	0.9977	1451	3	0.9979
10	2460	8	0.9968	1217	5	0.9959
11	857	1	0.9988	532	3	0.9944
12	308	0	1	156	0	1
13	34	0	1	29	0	1
14	12	0	1	6	0	1

Supplemental Table 3. Performance by matched path.

	Columbia					Cornell				
Matched Path	**True Positive**	**False Positive**	**False Negative**	**Sensitivity**	**PPV**	**True Positive**	**False Positive**	**False Negative**	**Sensitivity**	**PPV**
first	2772	56	37922	0.0681	0.9802	1585	81	38411	0.0396	0.9514
first,last	23531	438	15289	0.6062	0.9817	27745	303	9579	0.7434	0.9892
first,last,phone	25137	191	14638	0.632	0.9925	29824	302	9071	0.7668	0.99
first,last,phone,zip	22793	169	17164	0.5704	0.9926	24222	318	14640	0.6233	0.987
first,last,zip	27345	687	11349	0.7067	0.9755	28179	588	9180	0.7543	0.9796
first,phone	25718	281	13898	0.6492	0.9892	29919	358	8856	0.7716	0.9882
first,phone,zip	23311	262	16482	0.5858	0.9889	24422	447	14252	0.6315	0.982
first,zip	8073	554	31337	0.2048	0.9358	7677	835	29978	0.2039	0.9019
last	2237	104	38683	0.0547	0.9556	1156	82	39140	0.0287	0.9338
last,phone	12968	920	26167	0.3314	0.9338	6061	551	31221	0.1626	0.9167
last,phone,zip	12062	838	27342	0.3061	0.935	5582	542	32464	0.1467	0.9115
last,zip	5013	440	35327	0.1243	0.9193	3097	690	35540	0.0802	0.8178
phone	1393	936	37659	0.0357	0.5981	988	796	35771	0.0269	0.5538
phone,zip	1914	986	37278	0.0488	0.66	1506	738	36217	0.0399	0.6711

Supplemental Table 4. Relationship inference rules.

Person 1-2	Person 2-3	Person 1-3
Parent	Aunt/Uncle	Grandaunt/Granduncle
Parent	Child	Sibling
Parent	Grandchild	Child/Nephew/Niece
Parent	Grandparent	Great-grandparent
Parent	Nephew/Niece	Cousin
Parent	Parent	Grandparent
Parent	Sibling	Aunt/Uncle
Child	Aunt/Uncle	Sibling/Sibling-in-law
Child	Child	Grandchild
Child	Grandchild	Great-grandchild
Child	Grandparent	Parent/Parent-in-law
Child	Nephew/Niece	Grandchild/Grandchild-in-law
Child	Parent	Spouse
Child	Sibling	Child
Sibling	Aunt/Uncle	Aunt/Uncle
Sibling	Child	Nephew/Niece
Sibling	Grandchild	Grandnephew/Grandniece
Sibling	Grandparent	Grandparent
Sibling	Nephew/Niece	Child/Nephew/Niece
Sibling	Parent	Parent
Sibling	Sibling	Sibling
Aunt/Uncle	Aunt/Uncle	Grandaunt/Granduncle/Grandaunt-in-law/Granduncle-in-law
Aunt/Uncle	Child	Cousin
Aunt/Uncle	Grandchild	First cousin once removed
Aunt/Uncle	Grandparent	Great-grandparent/Great-grandparent-in-law
Aunt/Uncle	Nephew/Niece	Sibling/Cousin
Aunt/Uncle	Parent	Grandparent/Grandparent-in-law
Aunt/Uncle	Sibling	Parent/Aunt/Uncle
Grandchild	Aunt/Uncle	Child/Child-in-law
Grandchild	Child	Great-grandchild
Grandchild	Grandchild	Great-great-grandchild
Grandchild	Grandparent	Spouse
Grandchild	Nephew/Niece	Great-grandchild/Great-grandchild-in-law
Grandchild	Parent	Child/Child-in-law
Grandchild	Sibling	Grandchild
Grandparent	Aunt/Uncle	Great-grandaunt/Great-granduncle
Grandparent	Child	Parent/Aunt/Uncle
Grandparent	Grandchild	Sibling/Cousin
Grandparent	Grandparent	Great-great-grandparent
Grandparent	Nephew/Niece	First cousin once removed
Grandparent	Parent	Great-grandparent
Grandparent	Sibling	Grandaunt/Granduncle
Nephew/Niece	Aunt/Uncle	Sibling/Sibling-in-law
Nephew/Niece	Child	Grandnephew/Grandniece
Nephew/Niece	Grandchild	Great-grandnephew/Great-grandniece
Nephew/Niece	Grandparent	Parent/Parent-in-law
Nephew/Niece	Nephew/Niece	Grandnephew/Grandniece/Grandnephew-in-law/Grandniece-in-law
Nephew/Niece	Parent	Sibling/Sibling-in-law
Nephew/Niece	Sibling	Nephew/Niece/Nephew-in-law/Niece-in-law

Supplemental Table 5. Observational heritability of child abuse.

Observational Heritability of "Victim of Child Abuse" (V61.21) (N=1,142)

N Families Sampled	h_2^o	95% CI		p value	POSA	
100	0.69	0.46	0.96	1.66E-02	0.52	
200	0.76	0.45	0.97	6.89E-04	0.94	*
300	0.79	0.51	0.97	7.40E-06	1.00	*
400	0.84	0.53	0.98	4.00E-07	1.00	*
500	0.88	0.73	0.99	4.72E-09	1.00	*
600	0.90	0.73	1.00	4.19E-10	1.00	*

Excluding Siblings (N=1,089)

N Families Sampled	h_2^o	95% CI		p value	POSA	
100	0.77	0.61	1.00	4.24E-02	0.17	
200	0.67	0.46	0.92	3.36E-02	0.67	
300	0.68	0.44	0.99	8.74E-03	0.98	*
400	0.71	0.47	0.94	3.92E-04	1.0	*
500	0.72	0.50	1.00	8.24E-05	1.0	*
600	0.80	0.68	0.96	5.30E-06	1.0	*

Supplemental Table 6. 85 semi-manually created phenotypes.

Phenotype	Terminology	Codes	Modifier
Acne	ICD9	706.0, 706.1	
Alcoholism	ICD9	303	
Alzheimer's disease	ICD9	331	
Androgenic alopecia (females)	ICD9	704.00, 704.01, 704.02, 704.09	
Anorexia nervosa	ICD9	307.1	
Asthma	ICD9	493	
Attention deficit hyperactivity disorder	ICD9	314	
Autism	ICD9	299	
Bipolar disorder	ICD9	296.0, 296.4, 296.5, 296.6, 296.7, 296.80, 296.89	
Bladder cancer	ICD9	188	
Breast cancer	ICD9	174	
Bulimia nervosa	ICD9	307.51	
Cancer endocrine glands	ICD9	194	
Cancer Nervous system	ICD9	192, 200.50	
Cancer Nervous system age >15	ICD9	192, 200.50	Age=>15
Celiac disease	ICD9	579	
Cervical cancer	ICD9	180	
Cervix in situ cancer	ICD9	180	
Chronic obstructive pulmonary disease	ICD9	496	
Colon cancer	ICD9	153	
Colorectum cancer	ICD9	153, 154	
Coronary artery disease	ICD9	414.0, 414.2	
Coronary calcification	ICD9	414.4	
Corpus uteri cancer	ICD9	182	
Crohn's disease	ICD9	555.0, 555.1, 555.2, 555.9	
Depression	ICD9	311, 296.2, 296.3	
Discoid lupus erythematosus	ICD9	695.4	
Ectatic coronary lesions	ICD9	447.8	
Eczema (adults)	ICD9	691, 692	
Endometrial cancer	ICD9	182	
Epilepsy	ICD9	345	
Gallstone disease	ICD9	574	
Glaucoma	ICD9	365	
Graves' disease	ICD9	242	
Hangover (men)	ICD9	305	Sex=M
Hangover (women)	ICD9	305	Sex=F
Head and neck cancer	ICD9	195	
Heart disease	ICD9	410-414, 420-429	
Hypertension	ICD9	401-405	
Insomnia (current)	ICD9	307.41	
Insomnia (lifetime)	ICD9	307.42	
Irritable bowel syndrome (females)	ICD9	555.0, 555.1, 555.2, 555.9, 556	Gender=F
Leukemia	ICD9	208	
Leukemia age >15	ICD9	208	Age=>15
Lung cancer	ICD9	162	
Melanoma	ICD9	172	
Migraine	ICD9	346	
Nicotine dependence	ICD9	305.1	
Non-Hodking lymphoma	ICD9	202	
Obesity	ICD9	278	
Osteoarthritis (Distal interphalangeal joint - DIP)	ICD9	715.9	
Osteoarthritis (hip)	ICD9	715.15	

(continued)

Phenotype	Terminology	Codes	Modifier
Osteoarthritis (knee and hip)	ICD9	715.15, 715.16	
Osteoarthritis (knee)	ICD9	715.16	
Ovarian cancer	ICD9	183	
Pain	ICD9	338	
Pancreas cancer	ICD9	157	
Parkinson's disease	ICD9	332	
Periodontitis	ICD9	523	
Polycystic ovary syndrome	ICD9	256.4	
Prostate cancer	ICD9	185	
Psoriasis	ICD9	696	
Rectal and anal cancer	ICD9	154	
Rectum Cancer	ICD9	154	
Renal cancer	ICD9	189	
Rheumatoid arthritis	ICD9	714	
Rhinitis (children)	ICD9	477	
Rosacea	ICD9	695.3	
Schizophrenia	ICD9	295	
Sciatica	ICD9	724.3	
Skin cancer nonmelanoma	ICD9	173	
Stomach cancer	ICD9	151	
Stroke	ICD9	430, 431, 434, 436	
Systemic lupus erythematosus	ICD9	710	
Systemic lupus erythematosus (first-degree relative)	ICD9	710	Degree=1
Systemic lupus erythematosus (second-degree relative)	ICD9	710	Degree=2
Systemic lupus erythematosus (third-degree relative)	ICD9	710	Degree=3
Testicular cancer	ICD9	186	
Thyroid cancer	ICD9	193	
Tooth loss	ICD9	525.1	
Type-1 diabetes	ICD9	250.X1, 250.X3	
Type-2 diabetes	ICD9	250.X0, 250.X2	
Ulcerative colitis	ICD9	556	
Uterine cancer	ICD9	182	
Varicose veins	ICD9	454, 456	