Supplementary Materials for

Estimate of disease heritability using 7.4 million familial relationships inferred from electronic health records

Fernanda Polubriaginof, Rami Vanguri‡, Kayla Quinnies‡, Gillian M. Belbin‡, Alexandre Yahi, Hojjat Salmasian, Tal Lorberbaum, Victor Nwankwo, Li Li, Mark Shervey, Patricia Glowe, Iuliana Ionita-Laza, Mary Simmerling, George Hripcsak, Suzanne Bakken, David Goldstein, Krzysztof Kiryluk, Eimear E. Kenny, Joel Dudley, David K. Vawdrey†, Nicholas P. Tatonetti†

‡ These authors contributed equally, ordered in reverse alphabetical order

† Co-senior author

Correspondence to: nick.tatonetti@columbia.edu

This PDF file includes:

Materials and Methods

Figs. S1

Tables S1 to S5

Materials and Methods

The data for this study were obtained from the inpatient EHR used at the hospitals affiliated with three large academic medical centers in New York City: Columbia University Medical Center, Weill Cornell Medical College, and Mount Sinai Health System. Columbia University Medical Center and Weill Cornell Medical College 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. Similarly, we will refer to Mount Sinai Health System and its data as Mount Sinai.

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

This research was approved by the institutional review boards at the three 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 than 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 a parent of B, the opposite relationship is that B is a 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 the mother of patient B and patient B is the mother of patient C, then by inference we know that A is the grandmother of C and C is the grandchild of A. The rules used to perform these inferences are described in Table S4.

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 the 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 the gold standard to evaluate identified maternal relationships. True positive cases are when maternal relationships identified by our algorithm are also present in the EHR's mother-baby linkage table. False positive cases are when maternal relationships identified by our algorithm are discordant with the relationship available in the EHR's mother-baby linkage table. And lastly, false negative cases are 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). To assess if matches identified by different variables perform differently, we also computed sensitivity and PPV. We stratified the identified relationships by the number of variables used to match the emergency contact to a patient in a 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 were collected from existing sources for 1,524 individuals.

At Columbia, genotype data were available for 302 individuals. Data were 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. To select SNPs for kinship, minor allele frequency was filtered to >5%, and genotyping rate to 99% using PLINK (*15*). 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 identity by descent (IBD) 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%). The performance was evaluated by calculating PPV.

At Mount Sinai, we leveraged genome array data for 24,441 participants recruited to the Bio*Me* Biobank Program of The Charles Bronfman Institute for Personalized. Genotyped participants had a mean age 55.8 years, and approximately 61% are female. Participants self-identify as: Hispanic/Latino (45%), African American (31%), White/Caucasian (8%), Asian (6%), Mixed ancestry (6%), or Other (11%). To calculate genetic relatedness, we first merged Bio*Me* participants (N) genotyped either on the Illumina OmniExpress HumanCore (N=11,212)

and Multi-Ethnic Genotype Array v1.0 (N=10,467) platforms, retaining only the intersection of sites (n) between the two arrays (n=385,531). We subsequently removed palindromic sites (n=7,215 SNPs) and sites with a missingness rate > 1% (n=517) and a MAF < 5% (n=112,537) leaving a total of 112,537 SNPs. Of 21,679 Bio*Me* participants with genotype data, emergency contact information was available for 16,341, and in 1,222 cases both family members with relationship inferred by RIFTEHR were in Bio*Me*. Pairwise genetic relationships were estimated by Identity-by-State analysis with PLINK1.9 using the *–genome* flag. Inferred relationships from RIFTEHR were compared to pairwise genetic relationships to assess performance metrics using the "caret" package with R version 3.0.3. Pairs of patients with conflicting familial relationships were analyzed based on the closest relationship available. For example, if the same pair has two distinct relationships inferred based on their emergency contact information (e.g. parent and aunt/uncle), we consider the first-degree relationship to be correct (in this case, parent) for evaluation of the relationship against genetic data. Parent-offspring and sibling relationships groups were both expected to share ~50% genetic relatedness IBS (pi_hat mean 0.5, s.d. ± 0.1). We could distinguish between these two groups by examining the IBS measures at heterozygous (IBS1) and homozygous (IBS2) sites. Parent-offspring were defined as IBS1 > 0.75 and IBS2 < 0.25 (n=1087 pairs), full-siblings were defined as pairs that shared between 0.35 and 0.65 IBS1, and IBS2 > 0.15 and < 0.5 (n=502), monozygotic twins were defined as individuals sharing > 0.8 IBS2 (n=2). In each RIFTEHR group we calculated positive predictive values (PPV) based on how many predicted parent-offspring and siblings met this genetic criteria. Grandparental, avuncular and half-siblings are all expected to share ~25% genetic relatedness IBS (pi_hat mean 0.25, s.d. ± 0.05). We could not distinguish these groups any further, so calculated positive predictive values for each group based on how many total pairwise relationships met this criteria (n=976). We did not calculate PPV for cousins, grand-avuncular, great-grandparental, great-grand-avuncular, first cousin once removed relationships as the numbers of predicted relationships per group were low (n≤10). Finally, as negative control, we compared predicted

spousal relationships with low or no evidence of IBS sharing (pi_hat < 0.05, < 0.1 IBS1 and < 0.1 IBS2). The Bio*Me* Biobank Program (Institutional Review Board 07–0529) operates under a Mount Sinai Institutional Review Board-approved research protocol. All study participants provided written informed consent.

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 (e.g., laboratory tests such as hemoglobin A1c which is primarily used to measure the three-month average glucose concentration in plasma) as quantitative traits and diagnosis billing codes (ICD codes) as dichotomous traits. We extracted the most commonly performed laboratory tests and mapped them to LOINC codes so that they could be matched between institutions. Each patient may have multiple laboratory reports over time. To extract 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 laboratory report for the patient. 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." For ICD-9 codes, 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 ICD-9-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 with 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 from their diagnosis. Once this mapping was done, each phenotype was associated with one or multiple distinct CCS categories. We matched the 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.

For conditions recorded using ICD-10 codes, we use the hierarchy from ICD-10 to identify patients for the control group. Patients that did not have the same ICD-10 code as diagnosis nor any diagnosis that shared an ancestor code were considered controls.

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 S5).

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 hypothesized that repeated subsampling would be robust to biases introduced by extremely different ascertainment between families. We define the observational heritability, or h_o^2, 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 out of 155,811 and at Mount Sinai 25,438 families out of 187,473 contained conflicts. These families were excluded 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 1,404,671, 949,440 and 863,340 individuals for Columbia, Cornell, and Mount Sinai, 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. As the sample size increases towards the total number of available families the variance in heritability will decrease, but the estimate will be less robust to bias (Figure 3). This is because we are sampling without replacement. Based on our simulation studies we used sample sizes of 15 and 20% of the total number of families with at least one case. For those estimates that did not pass our quality control criteria at this level, we increased the number of families sampled to 45% 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 and report the median heritability with the 95% confidence interval.

3.2.2. Sample pedigree files. For each sampling, a set of N families is 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. To investigate the relative contribution of the environment to the studied phenotype, we used SOLAR to compute household effects. For this analysis, we assigned the mother ID as the household ID.

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 for a solution to be considered legitimate: (i) edge epsilon ($\in$e), any estimate within $\in$e of 1 or 0 is rejected; and (ii) noise epsilon ($\in$n), any estimate with implausibly low error is rejected (h2 error is less than $\in$n of the h2 estimate). These hyperparameters are set using simulated heritability data.

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 (a.k.a. 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 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. Or, 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 3I). We chose a minimum POSA score, $POSA_{lower}$ and the $\boldsymbol{\alpha_{POSA}}$ using simulations.

3.2.4. Aggregation of sampling results (computing h_o^2). For each sampling that passes quality control and meets the minimum POSA score, we compute the h_o^2 as the median. The median h_o^2 corresponds to a single run of SOLAR that has passed all quality control filters. We used the 95% confidence interval as the error of the h_o^2. We found that this error is closely related to the standard error reported by SOLAR (Figure 3). All raw heritability estimates that pass the initial quality control are made publicly available for reanalysis.

3.3. 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(*35*, *36*). 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 no 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.

3.4. Validation of accuracy and robustness of SOLARStrap using Simulated Traits

The scripts and data used in the following simulations are available publicly at https://github.com/tatonetti-lab/h2o.

3.4.1. Simulation of quantitative and dichotomous traits. We constructed a set of 4,195 families containing 14,690 individuals chosen from the families extracted from the EHR using RIFTEHR. Relationships and pedigree structures are heterogeneous across these families. We used the simqtl command from SOLAR to simulate quantitative traits with heritability values of 5%, 10%, ..., 90%, and 95% for this pedigree. Traits were simulated for 19 different heritability values in total. To generate quantitative traits, a threshold for the quantitative value was chosen for each of the 19 simulations so that the prevalence of the dichotomous, or binary trait, was 15%. The result of each simulation was a phenotype file (.phn) containing the family id, the individual id, and the quantitative or binary trait value.

3.4.2. Evaluation of simulated traits. We evaluated the quantitative and dichotomous traits by running SOLAR using the simulated phenotype files for each of the 19 different values for heritability. We summarize performance using the r-squared and run a test of significance.

3.4.3. Creating trait files for SOLARStrap. SOLARStrap is designed to use trait files that are similar to the phenotype files used by SOLAR but can contain more than one type of trait and more than 32,000 individuals (SOLAR's limit). We used a python script to combine the 19 heritability estimates into a single trait file.

3.4.4. Evaluation of the accuracy of SOLARStrap on quantitative traits. We ran SOLARStrap on each of the 19 simulated datasets. We repeated these runs using a different sampling size (argument nfam in SOLARStrap) between 100 and 700 increasing by 100. We selected the largest sample size (nfam=700) and evaluated the accuracy of SOLARStrap using r-squared and tested significance using regression analysis.

3.4.5. Evaluation of the accuracy of SOLARStrap on dichotomous traits. There are two scenarios when working with dichotomous traits. Either (1) the cases and controls are equally known. Meaning that each individual in the pedigree can be assigned to either being a case or control or (2) the cases are higher confidence than the controls. This latter case more closely resembles the scenario present in the electronic health records. Documentation of a disease in the EHR can be very indicative of the patient having the disease, but the absence of this documentation does not mean the patient does not have the disease. We evaluated the accuracy of SOLARStrap in both cases. For the former, we included all individuals in the pedigree, and for the latter, we excluded any families where there were no cases. In the latter case, we must also then assign a proband so that the estimate of heritability is not biased. We did this by randomly selected a single individual in each family as the "proband."

3.4.6. Evaluation of the robustness of SOLAR and SOLARStrap to missing data. To evaluate the robustness of SOLAR and SOLARStrap to missing data, we chose a single simulated trait (h^2=50%) and randomly changed individual phenotypes to NA. We evaluated removing 5%, 10%, ..., 55%, and 60% of the phenotype data.

3.4.7. Evaluation of the robustness of SOLAR and SOLARStrap to biased data (non-random missingness). To evaluate the robustness of SOLAR and SOLARStrap to biases, specifically non-random missingness, we used a beta distribution to assign a probability to each family of data being removed. By varying the beta distribution, we can control the amount of bias being introduced; higher beta values skew the distribution toward more extremes. We simulated non-random missingness using beta values of 0.001, 0.01, 0.1, 1.0, 10.0, and 100.0.

3.4.8. Evaluation of other measures of robustness and accuracy. Using the simulation results, we evaluated the effect of increasing the sample size (or the number of families being sampled in

each iteration when running SOLARStrap). We hypothesize that as the number of families approaches the number of available families the heritability estimate of SOLARStrap will converge to the heritability estimate of SOLAR. We expect that the number of families sampled would not have an effect on the heritability estimate produced by SOLAR or SOLARStrap. We evaluated this relationship using linear regression of the simulation results. One of the primary quality control metrics for SOLARStrap is the Proportion of Significant Attempts (or POSA). We evaluated the relationship between the POSA score (which ranges from 0 to 1) and the accuracy of the heritability estimates produced.

4. 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. First, we do not release data for any conditions where there are less than 100 individuals. Second, we do not release data for families containing more than five members. This will protect against identification through unique familial relationships situations. Third, we generate a new random map of patient identifiers for every individual trait. This will protect against the identification of an individual by looking for unique combinations of diseases. Unfortunately, this also will preclude the possibility of comorbidity analysis. Even with these additional limitations, our dataset constitutes one of the largest public releases of clinical data in history. All aggregate data and their corresponding statistics are released without obfuscation.

5. Computational and statistical software

Statistical analysis, data preparation, and figure creation were performed using Python 2.7. The python system environment is described fully in the supplemental materials. Relationship inferences were 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/.

6. Literature review

For validation purposes, we compared our heritability estimates to the ones reported in the most recent meta-analysis of twin correlations and heritability (MaTCH) (*7*). Using the ICD-10 hierarchy, we grouped our ICD codes to match the main chapters and subchapters reported in the MaTCH database. Since the meta-analysis grouped all traits into higher level traits, losing a lot of granularity, we also performed a 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.

Fig. S1. 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).

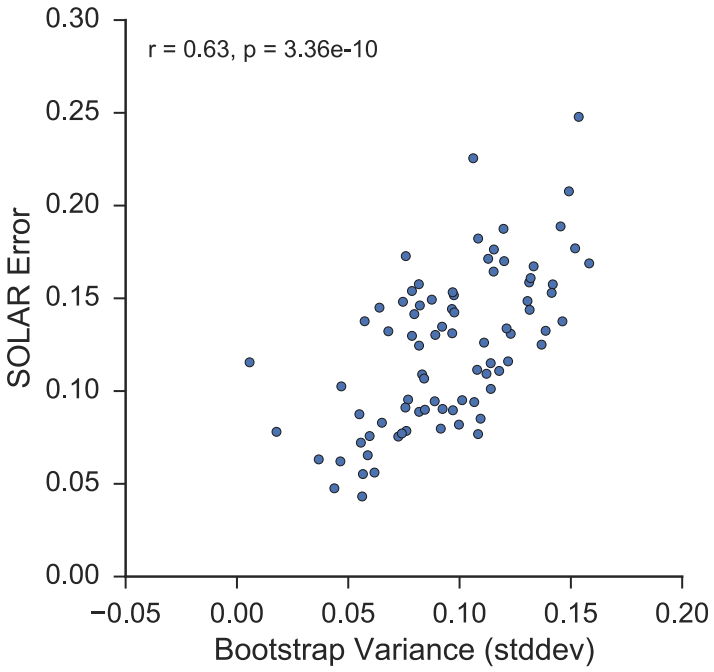

Table S1. Relationships by degree.

Degree of relationship	Relationship	N Columbia	N Cornell	N Mount Sinai
First	Child	482,308	298,136	252,584
	Parent	482,308	298,136	252,584
	Sibling	424,242	218,378	293,272
Second	Aunt/Uncle	185,822	65,410	75,404
	Nephew/Niece	185,822	65,410	75,404
	Grandparent	117,139	47,488	46,313
	Grandchild	117,139	47,488	46,313
Third	Cousin	148,806	37,370	27,994
	Grandaunt/Granduncle	96,675	31,764	36,069
	Grandnephew/Grandniece	96,675	31,764	36,069
	Great-grandchild	45,053	18,407	18,402
	Great-grandparent	45,053	18,407	18,402
Fourth	First cousin once removed	94,404	19,596	19,914
	Great-grandaunt/Great-granduncle	42,594	13,664	12,945
	Great-grandnephew/Great-grandniece	42,594	13,664	12,945
	Great-great-grandchild	17,854	7,531	6,348
	Great-great-grandparent	17,854	7,531	6,348
Other				
None	Child-in-law	0	278	0
	Parent-in-law	0	278	0
	Spouse	172,158	127,192	571,250
Unknown	Aunt/Uncle/Aunt-in-law/Uncle-in-law	13,220	5,234	45,950
	Child/Child-in-law	52,186	24,733	62,804
	Child/Nephew/Niece	31,818	8,078	96,925
	Grandaunt/Granduncle/Grandaunt-in-law/Granduncle-in-law	12,035	4,278	36,242
	Grandchild/Grandchild-in-law	12,876	4,578	32,781
	Grandnephew/Grandniece/Grandnephew-in-law/Grandniece-in-law	12,035	4,278	36,242
	Grandparent/Grandparent-in-law	12,876	4,578	32,781
	Great-grandchild/Great-grandchild-in-law	5,799	2,346	18,343
	Great-grandparent/Great-grandparent-in-law	5,799	2,346	18,343
	Nephew/Niece/Nephew-in-law/Niece-in-law	13,220	5,234	45,950
	Parent/Aunt/Uncle	31,818	8,078	96,925
	Parent/Parent-in-law	52,186	24,733	62,804
	Sibling/Cousin	41,270	9,142	88,956
	Sibling/Sibling-in-law	132,742	59,232	138,166

Table S2. 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

Table S3. 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

Table S4. 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

Table S5. 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	