ARTICLE OPEN ACCESS

Population genealogy resource shows evidence of familial clustering for Alzheimer disease

Lisa Anne Cannon-Albright, PhD, Sue Dintelman, MS, Tim Maness, BS, Johni Cerny, BS, Alun Thomas, PhD, Steven Backus, BS, James Michael Farnham, BS, Craig Carl Teerlink, PhD, Jorge Contreras, JD, John S.K. Kauwe, PhD, and Laurence J. Meyer, MD, PhD

Correspondence
Dr. Cannon-Albright
lisa.albright@utah.edu

Neurol Genet 2018;4:e249. doi:10.1212/NXG.0000000000000249

Abstract

Objective

To show the potential of a resource consisting of a genealogy of the US record linked to National Veterans Health Administration (VHA) patient data for investigation of the genetic contribution to health-related phenotypes, we present an analysis of familial clustering of VHA patients diagnosed with Alzheimer disease (AD).

Methods

Patients with AD were identified by the *International Classification of Diseases* code. The Genealogical Index of Familiality method was used to compare the average relatedness of VHA patients with AD with expected relatedness. Relative risks for AD were estimated in first- to fifth- degree relatives of patients with AD using population rates for AD.

Results

Evidence for significant excess relatedness and significantly elevated risks for AD in relatives was observed; multiple pedigrees with a significant excess of VHA patients with AD were identified.

Conclusions

This analysis of AD shows the nascent power of the US Veterans Genealogy Resource, in early stages, to provide evidence for familial clustering of multiple phenotypes, and shows the utility of this VHA genealogic resource for future genetic studies.

From the Genetic Epidemiology Program (L.A.C.-A., A.T., S.B., J.M.F., C.C.T.), Department of Internal Medicine, University of Utah School of Medicine; George E. Wahlen Department of Veterans Affairs Medical Center (L.A.C.-A., L.J.M.); Pleiades Software Development (S.D., T.M.), Inc, Salt Lake City; Lineages (J.C.), Draper; SJ Quinney College of Law (J.C.), University of Utah; Department of Biology (J.S.K.K.), Brigham Young University, Provo; Department of Dermatology (L.J.M.), University of Utah School of Medicine, Salt Lake City; and Department of Veterans Affairs (L.J.M.), Washington DC.

Funding information and disclosures are provided at the end of the article. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/NG.

The Article Processing Charge was funded by VHA.

Glossary

AD = Alzheimer disease; **CI** = confidence interval; **GIF** = Genealogical Index of Familiality; ***ICD*** = *International Classification of Diseases*; **PTSD** = posttraumatic stress disorder; **RR** = relative risk; **TBI** = traumatic brain injury; **UPDB** = Utah Population Database; **VHA** = Veterans Health Administration; **VISN** = Veterans Integrated Service Network.

The US Veterans Genealogy Project links a genealogy of the United States with medical data for Veterans who use the Veterans Health Administration (VHA) system. While still in its infancy, at a current size of 63 million individuals with genealogy data linked to 810,632 VHA patients, it is already sufficiently large for investigation of familial clustering for many health-related phenotypes.[1] With a resource that allows identification of individuals with a phenotype of interest, and for which biological relationships are known, it is possible to test for an underlying genetic predisposition.[2] This resource provides a unique opportunity to explore evidence for many phenotypes and to identify a rich resource of extended high-risk pedigrees.

Using this unique VHA resource still under creation, we present analysis of close and distant relationships among individuals diagnosed with Alzheimer disease (AD). This analysis of familial clustering in a population of US Veterans shows significant evidence for excess relatedness, significantly elevated risks in relatives, and identifies multiple extended high-risk pedigrees, confirming evidence supporting a genetic contribution to AD and displaying the potential of a powerful new national resource for predisposition gene identification.

Methods

US veterans genealogy

Genealogic data for over 63 million individuals gathered from public sources have been linked to a US genealogy. This resource is currently based on collected genealogy data that to date focused on Alaska, Arizona, Colorado, Idaho, Hawaii, Kansas, Montana, Nebraska, Nevada, New Mexico, North Dakota, South Dakota, Oklahoma, Oregon, Utah, Washington, Wyoming, and Massachusetts. Original sources used to compile the genealogy data include vital records, church records, censuses from 1850 to 1940, published genealogies, cemetery records, including gravestone inscriptions, family trees shared publicly online, oral history, and a variety of other sources used by genealogists to compile family trees. The demographic data for over 11 million Veterans using the VHA System was record-linked to this US genealogy data using GenMergeDB (pleiades-software.com), which has been used to create, and link records to, multiple genealogic resources for decades.[1] Over 810,000 VHA patients were record-linked to a unique individual in the genealogy using name, birthdate, and relationship data. After record linking was accomplished, no individual identifying data were used. The most important genealogic data for individuals are that for ancestors. We selected those VHA patients who linked to good ancestral data to allow more precise matching of controls; ancestral data allow us to identify more distantly related individuals in the same generation. Of the 810,632 VHA patients with linked genealogy data, 184,658 patients have genealogy data for at least 8 of their immediate ancestors, including at least both parents, all 4 grandparents, and at least 2 great grandparents (many patients have much more genealogy data). These VHA patients with at least 8 of their immediate ancestors were analyzed here.

Standard protocol approvals, registrations, and patient consents

Access to health data for the (unidentified) VHA patients with linked genealogy was approved by the University of Utah and Salt Lake Veterans Affairs Institutional Review Board, and approval was obtained from an oversight committee for the VHA resource.

Although the construction of the US genealogy to date has focused on genealogy sources with life events in the Western states, and the resource represents less than 25% of the final US genealogy to be created, VHA patients born in every state have been identified. Among the 810,632 Veterans who link to the genealogy, there are patients identified in all the 18 VHA Veterans Integrated Service Networks (VISNs or regions) across the United States. Of the 46% of the 810,632 linked VHA patients who have VISN data available, the largest numbers of linked VHA patients were in VISN 16 (South Central: 29,907), VISN 8 (Sunshine: 28,299), VISN 23 (Midwest: 25,618), and VISN 19 (Rocky Mountain: 23,291), and the smallest numbers of linked patients were in VISN 5 (Capitol: 7,740), VISN 2 (Upstate New York: 7,813), and VISN 3 (New York, New Jersey: 8,876). Among the 810,632 Veterans who link to the genealogy, there are 154,213 female patients (19%); among the 184,658 patients with good ancestral data, 15% are female. A wide age range of VHA patients linked to genealogy data, with birth years ranging from the early 1900s to the 1990s. The birth year distribution of the 810,632 VHA patients who linked to any genealogy differed slightly from that of the 184,658 VHA patients who had deeper ancestral genealogy data. Among the 810,632 linked VHA patients, 11% were born before 1911, compared with 14% of the 184,658 VHA patients who linked to ancestral data; 16% of all 810,632 VHA patients who linked to genealogy data were born in the 1960s to the 1990s, compared with 9% of the 184,658 VHA patients with ancestral data. These differences might be expected, given that males have higher record linking rates than females because of fewer name changes and that individuals born less recently can be expected to have more descendants.

VHA patients diagnosed with AD

The VHA has used an electronic medical record system at most VHA medical centers for inpatient and outpatient care since 1994. These records provide a rich source of phenotype data on the 11 million Veterans who use the system. *International Classification of Diseases (ICD) Revision 9* coding was used to identify patients with AD (331.0).

Genealogical Index of Familiality

The Genealogical Index of Familiality (GIF) test is a well-established method for testing for excess relatedness. It was developed for use with the Utah Population Database (UPDB), the first US genealogic resource used in research,[2,3] and has previously been used to establish evidence for many disease phenotypes, e.g., all cancer,[4,5] asthma mortality,[6] rotator cuff disease,[7] lumbar disc disease,[8] Alzheimer mortality,[9] and prostate cancer,[10] among others. A similar method has been used to establish evidence for familial clustering for a variety of medical conditions in the Icelandic Genealogy resource.[11]

The GIF statistic is a measure of the average pairwise relatedness for a set of individuals, for example, all VHA patients with AD. The pairwise relatedness is measured using the Malécot coefficient of kinship,[12] which is computed from genealogy information to estimate genetic relatedness for a pair of individuals. The coefficient of kinship estimates the probability that 2 alleles at a locus are identical by descent (inherited from a common ancestor) in a pair of individuals. All possible paths of relatedness are considered in the calculation. Most pairwise relationships in a large population-based genealogy are genetic distance = 0 (unrelated). For related pairs, the genetic distance increases with genetic distance, for example, for parent and offspring = 1, for siblings or for grandparent/grandchild = 2, for avunculars = 3, for first cousins = 4, for second cousins = 6, and, similarly, for more distant relationships. The GIF statistic is multiplied by 10^5 for ease of presentation.

The GIF test compares the average pairwise relatedness of a group of individuals to the expected average relatedness, which is estimated for a group of similar individuals in the population. The expected average pairwise relatedness for a set of VHA patients can be estimated for a randomly selected set of matched controls for the cases from the population of all VHA patients with linked genealogy data; controls are matched to cases for sex and 5-year birth year cohort. To estimate the mean expected pairwise relatedness, 1,000 sets of matched VHA controls were randomly selected and analyzed. The empirical significance of the GIF test was obtained by comparing the case GIF statistic to the distribution of the 1,000 control GIF statistics. This comparison of average pairwise relatedness tests whether the VHA patients with AD have significantly higher relatedness than expected in the VHA population. The GIF test does not allow determination of whether the familial clustering observed is due to environmental factors, genetic factors, or some combination. To consider whether familial clustering might primarily be due to shared exposures or behavior among close relatives, rather than shared genetics, the GIF method includes a distant relatedness test (dGIF). The dGIF test is performed as for the GIF test, but all relationships closer than third degree are ignored. Thus, the dGIF test ignores relationships most affected by shared environment or behavior and tests for the presence of excess distant relatedness only. Significant evidence for excess distant relatedness is strongly suggestive of a genetic contribution. The GIF statistic summarizes average pairwise relatedness in a single measure.

The contribution to the GIF statistic can be quantified separately for the different genetic distances observed among pairs of cases and controls (figure 1). The genetic distance measure represents, for example, 1 for parent/offspring, 2 for siblings or grandparent/grandchild, 3 for avunculars or similar, 4 for first cousins or similar, 6 for second cousins or similar, and so forth.

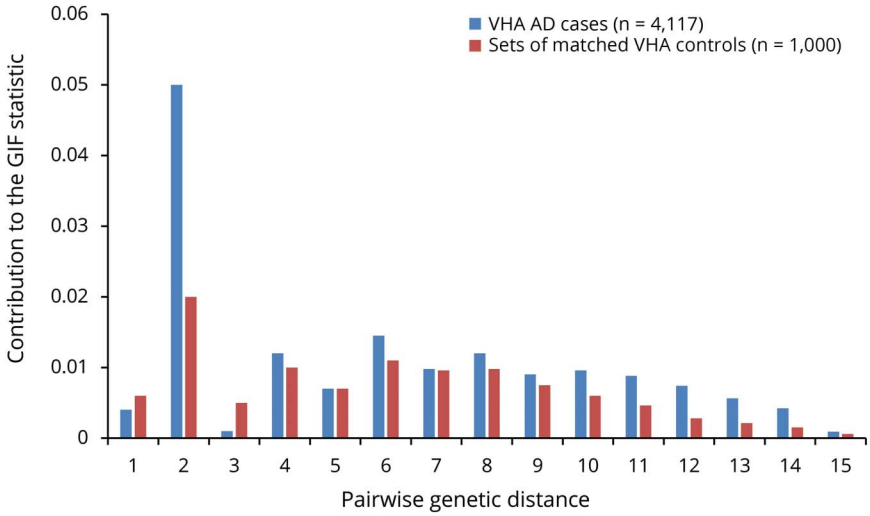
RRs in relatives

The estimation of relative risk (RR) in relatives provides a more traditional mechanism for identifying evidence for a genetic contribution. A genetic contribution to a phenotype is supported when both close and distant relatives show evidence of elevated risk. First-degree relatives include parents, siblings, or offspring; second-degree relatives are the first-degree relatives of first-degree relatives (e.g., uncle, grandmother); third-degree relatives are the first-degree relatives of second-degree relatives (e.g., first cousin, great grandchild), and so forth. RRs were estimated for first- to fifth-degree relatives of VHA patients diagnosed with AD as follows; all relatives considered were also VHA patients. All 184,658 patients in the VHA genealogy with genealogy data including at least 8 of 14 immediate ancestors were assigned to one of 67 cohorts based on birth year (in 5 years groups) and sex. The cohort-specific rate of AD was estimated as the number of AD cases in each cohort divided by the total number of linked VHA patients in the cohort. Expected numbers of first-degree relatives with AD were estimated by counting the number of relatives, all of whom were VHA patients with genealogy data, by cohort (without duplication), multiplying by the rate of AD in each cohort, and summing over all cohorts. Observed numbers of AD cases among relatives were counted without duplication. RRs were estimated for each degree of relationship (= observed/expected AD patients); 95% confidence intervals (CIs) for the RR were calculated using standard methods.[13]

High-risk pedigrees

To identify high-risk pedigrees for AD, all relationships among all VHA patients with AD were analyzed. Consideration of all ancestral vectors allowed identification of clusters of related cases; the nearest common ancestor was identified for each independent cluster of related patients with AD. No completely overlapping clusters were considered, but some cases appeared in more than 1 cluster (or pedigree). For

Figure 1 Contribution to the GIF statistic by pairwise genetic distance for cases compared with the average for 1,000 sets of matched controls



a given founder of a cluster of related patients, the number of observed AD cases among the descendants of the founder who were VHA patients was counted. To estimate the expected number of patients with AD among the descendants, all linked VHA patients among the descendants were counted by cohort; the number of linked VHA patients in each cohort was multiplied by the cohort-specific rate for AD (estimated as described above) and summed over all cohorts. A comparison of the number of observed to expected linked AD cases among the descendants in each cluster (pedigree) was made; if a significant excess of AD patients was observed ($p < 0.05$), the pedigree was termed high-risk.

Results

In the VHA resource, 4,117 Veterans with genealogy for at least 8 immediate ancestors and who had an *ICD-9* code indicating AD were identified; 194 (5%) were female. Table 1 summarizes the results of the GIF analysis for AD. The GIF test summary includes the sample size (n), GIF statistic for cases (case GIF), mean GIF statistic for 1,000 sets of controls (mean control GIF), empirical significance for comparison of overall GIF (empirical GIF p), distant GIF statistic for cases (case dGIF), mean dGIF statistic for controls (mean control dGIF), and empirical significance for comparison of dGIF (empirical dGIF p). The average pairwise relatedness for the 4,117 patients with AD was higher than expected for the VHA patient population ($p < 0.001$). When relationships closer than third degree (first cousins) were ignored, the average pairwise relatedness of the patients with AD was still elevated over expected relatedness (dGIF $p < 0.001$).

Figure 1 shows the contribution to the GIF statistic by the pairwise genetic distance for cases compared with averages for the 1,000 sets of matched controls. The effect of some data censoring based on the nature of the VHA data available can be observed in figure 1. Data censoring is present because

Table 1 GIF analysis in the VHA genealogy resource

		Mean			Mean		
Disease (*ICD-9* code)	**n**	**Case GIF**	**Control GIF**	**Empirical GIF *p***	**Case dGIF**	**Control dGIF**	**Empirical dGIF *p***
AD (331.0)	4,117	0.23	0.15	**<0.001**	0.15	0.11	**<0.001**
Matched AD controls	4,117	0.15	0.15	0.449	0.10	0.11	0.720
Prostate cancer (185)	12,695	0.18	0.14	**<0.001**	0.12	0.10	**0.003**
Parkinson disease (332)	3,850	0.22	0.15	**0.001**	0.15	0.11	**0.002**
Random set of patients	5,000	0.09	0.13	1.000	0.06	0.09	1.000

Bold values indicate statistical significance.
Abbreviations: AD = Alzheimer disease; GIF = Genealogical Index of Familiality; *ICD* = *International Classification of Diseases*.

diagnostic data for VHA patients were available only from 1994 to present. Because medical diagnosis was only available for slightly more than 20 years (1994–2017), and because AD is typically diagnosed in older ages, AD-affected relatives who are in the same generation (e.g., first degree: siblings or third degree: cousins) are the most likely to be observed; affected relatives in different generations (which includes all second- and fourth-degree relatives, e.g., a grandparent and grandchild) are unlikely to be observed in this narrow window. As the resource increases in size and years of data, this censoring will be lessened.

To validate that control matching and selection represented the VHA population and to demonstrate the overall baseline relatedness of VHA patients, we randomly selected 1 set of matched controls for the 4,117 patients with AD (termed "matched AD controls" in table 1); we treated this set of randomly selected VHA patients as a set of "cases" and performed GIF analysis to determine whether this single set of controls differed from 1,000 sets of matched controls (selected to match the single set of matched AD controls). This original set of controls for patients with AD did not differ in expected relatedness from the set of 1,000 sets of matched controls (GIF $p = 0.449$, dGIF $p = 0.720$, table 1).

Using the same methods, we also performed GIF analyses for VHA patients diagnosed with 2 other common phenotypes (prostate cancer and Parkinson disease) for purposes of comparison to AD and for comparison of results for these phenotypes reported from other resources. Prostate cancer cases were identified with *ICD-9* code 185, and Parkinson disease cases were identified with *ICD-9* code 332. The overall GIF test and the distant dGIF test showed significant excess relatedness for both phenotypes (table 1). These analyses confirm previously published evidence of significant excess relatedness for both close and distant relationships for these 2 phenotypes from population-based genealogy resources in Utah,[4,5,14] Iceland[11] and Scandinavia.[15] In addition, to demonstrate that not all sets of VHA patients show greater than expected excess relatedness, we randomly selected 5,000 VHA patients with genealogy data and no associated phenotype. Results for the GIF analysis of this random set of patients indicate no excess relatedness (table 1).

RRs in relatives

Estimated RRs for AD among relatives of patients with AD who are also VHA patients are shown in table 2, which displays degree relatedness, total number of relatives among linked VHA patients (n), observed number of relatives with AD who were VHA patients (obs), expected number of relatives with AD who were VHA patients (exp), RR, significance (p value), and 95% CI for the RR (95% CI). RRs for AD were significantly elevated among first- (RR = 1.82) and fifth-degree relatives (RR = 1.22) of patients with AD who were VHA patients and were elevated (RR = 1.06), but not significantly ($p = 0.380$), among third-degree relatives. RR results for second- and fourth-degree relatives are affected by data censoring issues, as discussed previously. This is apparent when, for example, the different types of first-degree relatives are considered separately; 558 of the 882 first-degree relatives identified were siblings, whereas only 274 were children and 50 were parents. Because only 5% of the VHA AD patients were female, comparisons of effects by sex were not possible. For example, all the 41 affected sibling pairs observed were brothers, and both of the observed affected children of affected parents were sons.

High-risk pedigrees

Two hundred forty-five high-risk AD pedigrees were identified ($p < 0.05$), with at least 2 and up to 117 related VHA AD patients. Figure 2 shows an example high-risk AD pedigree identified in the VHA genealogy resource; only the descending lines to the VHA AD patients are shown. The pedigree includes 6 related VHA AD patients, only 1.2 AD cases were expected among the descendants who were VHA patients ($p = 0.0012$). The pedigree founder was born in Pennsylvania in the late 1700s and has almost 6,000 descendants in the genealogy; 67 descendants are VHA-linked patients.

Discussion

The VHA Genealogy Resource is a unique resource that continues to grow and improve. We used this partially constructed US genealogy of over 63 million individuals linked to almost 1 million patient records representing all VHA local areas to show the potential for genetic analyses, using AD, a complex disease, as an example. The Alzheimer's Association (2015) has reported that AD is the sixth leading cause of death in the United States. The annual cost of dementia in the United States has been estimated to be $215 billion in 2010 and is expected to double by 2040.[16] One study[17] projected 13.8 million people diagnosed with AD dementia by 2050 in the United States. Although old age is the primary risk factor for AD, a genetic contribution to AD predisposition is also well recognized.[18] Mutations in AβPP, Presenilin 1, and Presenilin 2 have been implicated in familial or early-onset AD; the APOE ε4 allele is a major genetic risk factor for AD; and other genetic risk factors involving lipid metabolism and immune function are recognized.[19–22]

Both traumatic brain injury (TBI) and posttraumatic stress disorder (PTSD) have been linked to an increased risk of AD and other dementias, and both are "signature injuries" of individuals serving in the Iraq and Afghanistan conflicts.[23,24] Although AD is recognized as an important public health issue, the association of these military-related injuries with AD makes it of particular importance among military health issues. The association of TBI and PTSD with an increased risk of AD suggests that it may be valuable to identify those military personnel at high risk of AD and to develop interventions that could limit the progression or onset of disease. The results of this analysis of the VHA population, in combination with similar studies, suggest that there is a genetic contribution to AD and that predisposition can already be

Table 2 Estimated RRs for AD among first-to fifth-degree relatives of patients with AD in the VHA genealogy resource

Degree relatedness	n	obs	exp	RR	p Value	95% CI
First degree	882	45	24.7	**1.82**	**0.0001**	**1.33–2.44**
Sibling	558	41	20.6	**1.99**	**$4.7e^{-5}$**	**1.43–2.70**
Brother	527	41	20.1	**2.04**	**$2.8e^{-5}$**	**1.46–2.77**
Parent	50	2	1.9	1.05	1.000	0.13–3.78
Child	274	2	2.2	0.90	0.617	0.11–3.25
Son	231	2	2.2	0.93	0.637	0.11–3.37
Second degree	633	2	7.7	0.26	0.018	0.03–0.94
Grandparent	11	0	0.4			
Grandchild	186	0	0.3	0.00	0.759	
Father's brother	28	0	1.1			
Mother's brother	27	1	1.2	0.86	0.674	0.02–4.77
Sister's son	157	1	2.3	0.43	0.322	
Brother's son	133	0	1.6	0.00	0.213	
Half-sibs	18	0	0.4	0.00	0.670	
Father's sister	2	0	0.1	0.00	0.931	
Mother's sister	3	0	0.03	0.00	0.967	
Brother's daughter	29	0	0.1	0.00	0.917	
Sister's daughter	39	0	0.2	0.00	0.820	
Third degree	1,226	37	34.9	1.06	0.380	0.75–1.46
First cousins	984	37	33.4	1.11	0.289	0.78–1.53
Fourth degree	1,896	35	41.9	0.84	0.161	0.58–1.16
Fifth degree	4,517	152	124.6	**1.22**	**0.016**	**1.03–1.43**

Bold values indicate statistical significance.
Abbreviations: AD = Alzheimer disease; CI = confidence interval; exp = expected number of relatives with AD who were patients with VHA; obs = observed number of relatives with AD who were patients with VHA; RR = relative risk.

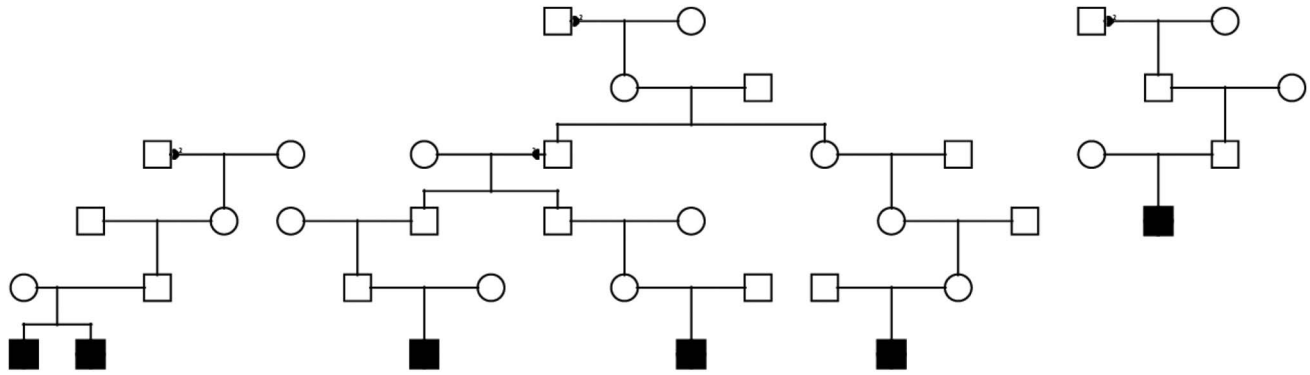
recognized through knowledge of family history of AD. As AD predisposition genes are identified, increased risk may also be recognized by screening.

Our analyses demonstrate significant evidence for excess familial clustering of VHA patients with AD compared with expected clustering in this population; significantly elevated RRs for AD in both close and distant relatives were observed, and many pedigrees with a significant excess of AD cases have been identified. This combined evidence confirms a heritable contribution to the observed familial clustering of AD. The pedigrees we identified suggest a highly elevated risk among some families. Such pedigrees are the ideal starting point for whole-genome sequence–based approaches to the identification and characterization of rare high-penetrance variants.

In addition to the evidence presented for excess relatedness of individuals diagnosed with AD, using the same methods, we have replicated published evidence for excess familial relatedness for 2 other common phenotypes to generalize validation of the resource. The 2 common disease phenotypes examined (prostate cancer and Parkinson disease) have previously been recognized to have a heritable component in other populations. Analysis of the VHA genealogy resource confirmed significant evidence for excess relatedness for both phenotypes; these results additionally validate the US VHA Genealogy Resource in terms of data quality and power for analysis of familial clustering.

There are limitations to this analysis. It is optimal to match controls based on all characteristics that might affect record linking or that are associated with the phenotype examined; we know that birth year, sex, and birth state (Utah or not) affect the overall relatedness of individuals in the similar UPDB genealogic resource that represents the Utah population.[2] For this analysis of the VHA resource, matching was performed only for birth year and sex. Although the rate of AD

Figure 2 Example high-risk AD pedigree identified in the VHA resource



Male founder has 2 marriages as does male grandson of the founder's first marriage; fully shaded are AD cases.

was similar (2%) in both the 810,632 VHA patients linking to any genealogy and the 184,658 VHA patients with at least 8 of their immediate ancestors, the VHA patients with at least 8 of their immediate ancestors had a slightly lower rate of females and represented a population born slightly later than all VHA patients who linked to genealogy. It is not clear what effect these slight differences might have had on the results. Historically, race data have not been stored for the majority of demographic records in the VHA system, and so was unavailable. In future, as more data become available, we propose to use data including birth state, VISN, occupational exposures, rank, and socioeconomic status, for example, for matching.

Data censoring is also an issue for this resource. VHA cases who fail to link to genealogy data are censored, as are diagnoses made outside the VHA system, or before 1994. This censoring of cases might affect the estimation of rates of AD; however, because rates were estimated for the entire linked population of VHA patients and were only used as relative comparisons, this censoring is not expected to affect results or tests of hypotheses. The overall rate of AD among all VHA patients with linked genealogy data was 2.2% (4,117/184,658). AD rates ranged from ~5% among male patients born before 1925 decreasing to 0.1% in male patients born in the 1960s, with rates of ~3% in females born before 1925 decreasing to 0.1% in female patients born in the 1950s.

In addition, genealogy data may not always represent biological relationships. However, such censoring is assumed to be independent of phenotype and equally affects both cases and controls. Finally, the data set is limited to primarily males, individuals who are part of groups who have shared genealogic data, and veterans who used the VHA system; this is true of both cases and controls, but may have affected results. The familial clustering methods presented are very robust to data censoring. Controls are VHA patients, matched for sex and birth year, and are required to have linkage to genealogy data of similar quality and quantity as cases. It is difficult to conceive of a mechanism by which significant excess relatedness would exhibit itself in the VHA resource in the absence of any true heritable component. Because of censoring, it is much more likely that the evidence of excess relatedness presented is conservatively estimated. The relatedness analysis of a set of matched controls considered as cases and the results for the 5,000 random VHA patients with no selected phenotype both demonstrated that the analysis method appropriately observes no evidence for familial clustering for these examples.

The methods presented are robust to misclassification of cases. False negatives (missing identification of true VHA AD cases) could result in failing to observe evidence for excess relatedness; this did not occur. False positives, even at a very high rate, could only affect this familial clustering analysis if the assignment of the incorrect diagnosis of AD in a VHA patient occurred more often among close and distant relatives of AD cases than among all VHA patients. This is unlikely, given the US-wide coverage of VHA patients represented and the distance of the genetic relationships providing evidence of excess clustering.

This VHA Genealogy Resource represents what we believe is already the largest genealogy linked to phenotype data based on its current size of over 63 million individuals and its linkage to medical data for over 810,000 VHA patients. This resource is still under construction; we estimate that the eventual size of this US genealogy will exceed 300 million individuals, with 40%–60% of the 11 million VHA patients with demographic data linked to genealogy. The analysis of the clustering of AD presented here is 1 example of the utility of the resource for genetic studies. The utility of this resource includes (1) demonstration of evidence for a genetic contribution to predisposition to many health-related phenotypes not commonly observed in other populations; (2) identification and study of high-risk pedigrees informative for predisposition gene identification; (3) estimation of family history–based risk of any disorder of interest, which may be widely applicable to the US

population; and (4) identification of both high- and low-risk individuals for any phenotype of interest for epidemiologic studies or clinical trials, among others. This resource may be uniquely powerful for analysis of phenotypes that are rarely observed and highly associated with military service and may have an underlying genetic contribution (e.g., PTSD).

Clearly, this resource will improve in 2 distinct ways. Expansion of the genealogy data to all states is in progress. Second, there is enormous potential to refine the phenotypes analyzed. The use of *ICD-9* diagnostic coding to identify individuals with a phenotype is not optimal; such coding has other purposes than research and may misrepresent the phenotype in both directions (false positives and false negatives). Future analysis of this resource will dictate more refined phenotype definitions. The Electronic Medical Records and Genomics Network has demonstrated the benefit of more robust analysis using multiple components of the medical record.[25] Use of natural language processing algorithms from text data in medical notes may allow better identification of phenotypes on a large scale. The VHA is well positioned to take advantage of these methods, specifically using the VHA Informatics and Computing Infrastructure resource.

In conjunction with the recent Million Veterans Program sponsored by the VHA, which is collecting and storing DNA and demographic and risk data for 1 million VHA patients, extremely powerful genetic studies will soon be possible. With the future additions of genotypes, exposure data, and other data that are envisioned, this VHA genealogy/phenotype resource will allow informative genetic studies that include relationship data on an extremely large scale, including gene by environment analyses, and analysis of other medical conditions related to service, which cannot be studied in most populations.

An initial genetic analysis of AD has been presented using a powerful new and growing national resource linking genealogy and medical data. AD was selected as the example phenotype, given that it is a major health issue in the United States and the world and may be an important issue for the VHA in light of reported associations of AD with trauma. The analyses confirmed evidence for an inherited component to AD risk, identified a current resource of high-risk pedigrees that could be used for predisposition gene/variant identification, and confirmed the power and utility of this VHA resource for genetic studies of complex human disease.

Author contributions

L.A. Cannon-Albright: study concept, resource creation, study design, analysis, and manuscript preparation. S. Dintelman and T. Maness: genealogy data resource design and construction and record linking. J. Cerny: genealogy data collection and organization. A. Thomas: refinement of analysis methods and tools. S. Backus: creation of genealogy/phenotype resource and analysis tools. J.M. Farnham: refinement of analysis tools. C.C. Teerlink, J. Contreras, and J.S.K. Kauwe: critical refinement of the manuscript. L.J. Meyer: resource concept and manuscript preparation.

Study funding

This material is based on work supported in part by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, and RF1 AG054052 (J.S.K.K, PI).

Disclosure

L.A. Cannon-Albright has received research support from the NIH, US Department of Veterans Affairs, US Department of Defense, and Intermountain Research and Medical Foundation. S. Dintelman is or has been employed by Pleiades Software Development, Inc and has received research support from the Department of Veterans Affairs. T. Maness is or has been employed by Pleiades Software Development, Inc and has received research support from the Department of Veterans Affairs and Veterans Health Administration Office of Research and Development. J. Cerny is or has been employed by Lineages, Inc. and has received research support from the Department of Veterans Affairs. A. Thomas reports no disclosures. S. Backus has received research support from the US Department of Health and Human Services, Veterans Affairs Salt Lake City Health Care System, Army Medical Research Acquisition, and NIH. J.M. Farnham reports no disclosures. C.C. Teerlink has received research support from the NIH, Department of Veterans Affairs, and Department of Defense. J. Contreras has received funding for travel and/or speaker honoraria from American University, Stanford University, Arizona State University, University of Texas, Intel Corp, Standards Engineering Society, Centre for International Governance Innovation, Harvard University, Tilburg University (Netherlands), American Antitrust Institute, Jindal Global University, Practicing Law Institute, University of California Berkeley, University of Nevada Las Vegas, Brigham Young University, National Law University—Delhi, India, Leiden University, Netherlands, National Law School University, Hispanic National Bar Association, Creative Commons, Saint Louis University, Hunan University (China), RIETI (Japan), Japan Fair Trade Commission, and Duke University; receives or has received publishing royalties related to the Cambridge Handbook of Technical Standardization Law, Vol. 1, Cambridge Univ. Press 2017; has served as a consultant for Internet Society/IETF, International SAE Consortium Ltd, Hillcrest Laboratories, Inc., Natl. Assn. of State Securities Agents, IPBridge, William Hagmaier, BLU Products, Wilson Electronics, Industry Pharmacogenomic Working Group; has received research support from the NIH, Arizona State University, University of Texas, International Development Research Centre (IDRC), Centre for International Governance Innovation, and Huntsman Cancer Institute and Foundation; and has participated in legal proceedings for the Internal Revenue Service and IPBridge. J. S.K. Kauwe serves or has served on scientific advisory boards of ADx Health Care; has served on the editorial board of *Alzheimer's & Dementia*; holds patents related to systems, assays,

and methods for determining risk factors for Alzheimer's disease; and serves or has served as a consultant for Genoma LLC. Laurence L.J. Meyer reports no disclosures. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/NG.

Received December 4, 2017. Accepted in final form May 24, 2018.

References

1. Cannon-Albright LA, Dintelman S, Maness T, Backus S, Thomas A, Meyer LJ. Creation of a national resource with linked genealogy and phenotypic data: the Veterans Genealogy Project. Genet Med 2013;15:541–547.
2. Cannon-Albright LA. Utah family-based analysis: past, present and future. Hum Hered 2008;65:209–220.
3. Skolnick M. The Utah Geneological Database: a resource for genetic epidemiology. In: Banbury Report No 4: Cancer Incidence in Defined Populations. Cold Spring Harbor: Cold Spring Harbor Laboratories; 1980.
4. Cannon L, Bishop DT, Skolnick MH, Hunt S, Lyon JL, Smart CR. Genetic epidemiology of prostate cancer in the Utah Mormon genealogy. Cancer Surv 1982;1:48–69.
5. Cannon-Albright LA, Thomas A, Goldgar DE, et al. Familiality of cancer in Utah. Cancer Res 1994;54:2378–2385.
6. Teerlink CC, Hegewald MJ, Cannon-Albright LA. A genealogical assessment of heritable predisposition to asthma mortality. Am J Respir Crit Care Med 2007;176:865–870.
7. Tashjian RZ, Farnham JM, Albright FS, Teerlink CC, Cannon-Albright LA. Evidence for an inherited predisposition contributing to the risk for rotator cuff disease. J Bone Joint Surg Am 2009;91:1136–1142.
8. Patel AA, Spiker WR, Daubs M, Brodke D, Cannon-Albright LA. Evidence for an inherited predisposition to lumbar disc disease. J Bone Joint Surg Am 2011;93:225–229.
9. Kauwe JS, Ridge PG, Foster NL, Cannon-Albright LA. Strong evidence for a genetic contribution to late-onset Alzheimer's disease mortality: a population-based study. PLoS One 2013;8:e77087.
10. Nelson Q, Agarwal N, Stephenson R, Cannon-Albright LA. A population-based analysis of clustering identifies a strong genetic contribution to lethal prostate cancer. Front Genet 2013;4:152.
11. Sveinbjornsdottir S, Hicks AA, Jonsson T, et al. Familial aggregation of Parkinsons's disease in Iceland. NEJM 2000;343:1765–1770.
12. Malecot G. Les Mathematiques de l'heredite. Paris: Masson & Cie; 1948.
13. Agresti A. Categorical Data Analysis. New York: Wiley; 1990.
14. Savica R, Cannon-Albright LA, Pulst S. Familial aggregation of Parkinson disease in Utah: a population-based analysis using death certificates. Neurol Genet 2016;2:e65.
15. Hemminki K, Dong C. Familial prostate cancer from the family-cancer database. Eur J Cancer 2000;36:229–234.
16. Delavande A, Hurd MD, Martorell P, Langa KM. Dementia and out-of-pocket spending on health care services. Alzheimers Dement 2013;9:19–29.
17. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. Neurology 2013;80:1778–1783.
18. Ridge PG, Hoyt KB, Boehme K, et al. "Assessment of the genetic variance of late-onset Alzheimer's disease." Neurobiol Aging 2016;41:213–220.
19. Harold D, Abraham R, Hollingworth P, et al. "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease." Nat Genet 2009;41:1088–1093.
20. Hollingworth P, Harold D, Sims R, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet 2011;43:429–435.
21. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease." Nat Genet 2013;45:1452–1458.
22. Ridge PG, Ebbert MT, Kauwe JS. Genetics of Alzheimer's disease. Biomed Res Int 2013;2013:254954.
23. Yaffe K, Vittinghoff E, Lindquist K, et al. Posttraumatic stress disorder and risk of dementia among US veterans. Arch Gen Psychiatry 2010;67:608–613.
24. Plassman BL, Havlik RJ, Steffens DC, et al. Documented head injury in early adulthood and risk of Alzheimer's disease and other dementias. Neurology 2000;55:1158–1166.
25. Gottesman O, Kuivaniemi H, Tromp G, et al. The electronic medical records and genomics (eMERGE) Network; past, present, and future. Genet Med 2013;15:761–771.