ORIGINAL RESEARCH ARTICLE | Genetics inMedicine

Creation of a national resource with linked genealogy and phenotypic data: the Veterans Genealogy Project

Lisa A. Cannon-Albright, PhD[1,2], Sue Dintelman, MS[3], Tim Maness, BS[3], Steve Backus, BS[1], Alun Thomas, PhD[1] and Laurence J. Meyer, MD, PhD[2,4]

Purpose: Creation of a genealogy of the United States and its ancestral populations is under way. When complete, this US genealogy will be record linked to the National Veteran's Health Administration medical data representing more than 8 million US veterans.

Methods: Genealogical data are gathered from public sources, primarily the Internet. Record linking using data from relatives is accomplished to integrate multiple data sources and then to link genealogical data to the veteran's demographic data.

Results: This resource currently includes genealogy for more than 22 million individuals representing the Intermountain West and the East Coast. The demographic data for more than 40,000 veteran patients using Veterans Hospital Administration services in Utah and Massachusetts have already been record linked.

Conclusion: The resource is only in its second year of creation and already represents the largest such combination of genealogy and medical data in the world. The data sources, the creation of the genealogy, record-linking methods and results, proposed genetic analyses, and future directions are discussed.

Genet Med 2013:15(7):541–547

Key Words: electronic medical record; genealogy; record linking; veterans

INTRODUCTION

The Utah Population Database and the deCode Genetics Icelandic genealogy are well-recognized resources that combine genealogic data with phenotypic data for defined populations (ref.1; http://www.decode.com/research). These two resources have proven extremely valuable in understanding the genetic contribution to complex phenotypes[2–10] and in predisposition gene identification.[11–15] The Veterans Genealogy Project reported here proposes extension of this general concept by creation of a genealogy/biomedical resource for the United States and its immigrant founders. The extensive Veterans Administration (VA) electronic medical record system, consisting of medical data for more than 8 million of the national VA population of retired servicemen and -women who choose to use this medical service, will be record linked to this genealogy.

The VA genealogy project began in 2010, and this resource is still under construction. The resource currently includes more than 22 million individuals in a genealogy that largely represents Utah, Massachusetts, and their surrounding states. The demographic data for patients using VA facilities in either Utah or Massachusetts have been record linked to this VA genealogy for 41,290 VA patients. This initial VA biomedical resource consisting of two states is already a powerful and informative database; it allows study of the familial/genetic clustering of phenotypes represented within the genealogy data (e.g., longevity) and phenotypes represented in the medical data that relate more significantly to the VA population (e.g., posttraumatic stress syndrome). It is the largest such resource in the world and will become more informative as new genealogy data for other states are added and more VA patients linked. Here we describe the resource, how it is being created, how it can be used to enhance knowledge of the genetic contribution to health, and future directions.

MATERIALS AND METHODS

US genealogy data

The data used to build the US genealogy have been harvested from publicly available sources on the Internet. Because genealogy is a popular hobby, many websites contain genealogy data. We selected Utah and Massachusetts to begin to build this first phase of the US genealogy because a large amount of genealogy data is available for these two states. This is primarily due to the interest of members of the Church of Jesus Christ of the Latter-Day Saints (Mormons or LDS) in genealogy research, as well as the interest of individuals in the United States to trace their lineage back to pioneers on the *Mayflower*. We identified and collected data from genealogy sites that contain information in standard formats. We have collected more than 70,000 genealogy data sources; sources with events in Utah, Massachusetts, or surrounding states have been integrated here. Such publically available genealogy data can represent thousands of individuals in a family and are typically created over many decades or centuries, with families building and sharing genealogy data.

The Veterans Genealogy Project is in its third year of creation. The first year was spent building the database using sources for Utah genealogy. A number of these sources contained life

[1]Division of Genetic Epidemiology, Department of Medicine, University of Utah School of Medicine, Salt Lake City, Utah, USA; [2]Research and Development Service, George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, Utah, USA; [3]Pleiades Software Development, Inc, Salt Lake City, Utah, USA; [4]Department of Dermatology, University of Utah School of Medicine, Salt Lake City, Utah, USA. Correspondence: Lisa A. Cannon-Albright (lisa.albright@utah.edu)

Submitted 23 July 2012; accepted 15 November 2012; advance online publication 17 January 2013. doi:10.1038/gim.2012.171

events for Massachusetts. Extensive searches for Massachusetts genealogy sources have just begun, and these are not yet well represented in the genealogy. The expansion of Massachusetts data, followed by other US states, is ongoing.

Typically, genealogy data sources represent one of two scenarios: a summary of all the ancestors of an individual or a summary of all the descendants of an individual. To create a genealogy, these different data sources must be curated and combined. Multiple genealogy sources may reference the same individuals, sometimes with different data (e.g., number of spouses or children, or with variations in names). A more complete picture of an individual can be established by finding, and combining, information for this same individual from different sources. The resulting combined records are called composite records. The result of record linking to produce a genealogy is a set of composite individuals. When combining information from multiple genealogy sources, the record-linking system attempts to answer the question "Do these two records represent the same individual?" The methods used to best answer this question are discussed below.

Record linking

Both creation of the US genealogy and combining the VA patient demographic data with this genealogy are record-linking tasks. The record-linking system GenMergeDB (http://www.pleiades-software.com), which has been developed and tested on multiple genealogic resources over the past three decades, was used to build this resource. The details of the methodology are provided below.

Scoring links

GenMergeDB is based on probabilistic record linking.[16,17] Using this method, two records are compared for common fields (names, dates, and places); if the fields match, a positive score is assigned to that field; if the fields do not match, a negative score is assigned to that field. The scores associated with the outcome of each field comparison are summed, and if the result is over a threshold, the two records are considered matched. The score or weight associated with each matched field is computed using the frequency of the field value in the data sets being linked, so that commonly observed values are discounted and rarely observed values have enhanced value:

$$W_{\text{field}} = \log_2(p/a_i)$$

where p is the population size and a_i is the absolute frequency of the value i in the population.

In record linking, the fields with the most discriminating power are names. In historical data, there are many reasons why names might differ in different data sources. There are recording errors, transcription errors, and actual name variations and changes (e.g., anglicizing names when immigrating or name changes at marriage). To match only records with exact name matches would eliminate many good links. Much of the work in record linking focuses on string-matching algorithms, allowing inexact matching. GenMergeDB implements several string comparison algorithms that allow partial matches on strings when the changes are phonetic (e.g., Albertson, Albertsen) or caused by transcription errors (e.g., Thompson, Tjompson). Three algorithms used in GenMergeDB are the Utah Phonetic Transducer (V. Wesley, unpublished data) NYIIS,[18] and the Jaro-Winkler String comparator.[19] Names are compared using all methods, and the best score is used. Differences in dates and places also exist in historical data, and similarly, positive scores can be assigned to dates and places that do not match exactly.

Building the genealogy

The record-linking process for genealogy construction consists of three steps. In step 1, all the records are binned using some criteria (e.g., a coded value for the last and first names). Then, each record in a bin is compared with every other record in that bin, and pairs that score over a threshold are added to a set of clusters. The next two steps are unique to the GenMergeDB record-linking system. These steps allow linking of low-information relatives by focusing on small family-level linking problems. Step 2 uses the initial set of clustered records to revisit family members. During this pass, it is possible to link two records together that have extreme variations or little information based on matching data for relatives. Step 3 is similar but focuses on linked individuals whose parents did not match. Pairs of records in this category may actually have different mothers and fathers, indicating a problem with the pedigree, but can also be the result of large name or date variations. By using just the records for the parents, it is possible to apply much different scoring thresholds than are used in the previous two steps. This use of family data to augment the individual score allows confident matches to be made even for low-scoring individuals.

The resulting genealogy contains records composited from all sources. Because of the overlap in individuals across sources, this genealogy is largely interconnected and contains deep pedigrees, some extending back more than 15 generations from the present. Because published genealogy data are based largely on extensive family research of historical records, there is a high expectation of good-quality data; however, unacknowledged consanguinity, misattributed paternity, or the inclusion of "social" rather than "biological" relationships may be included. Such genealogical errors are expected to be infrequent, and we assume they are represented randomly across the resource.

Validity of the genealogy is verified in the following ways:

1. Comparison of key family information between original sources and the final linked genealogy. Aggregate demographic information is created for the set of pedigrees before linking. This information is monitored during the linking process to ensure linking criteria and cutoff thresholds remain correct.[20] Comparing the number of spouses per marriage, number of children per sibship, average time span from birth of youngest child to birth of oldest child, age difference of spouses at marriage, and other measures in the final linked genealogy to the

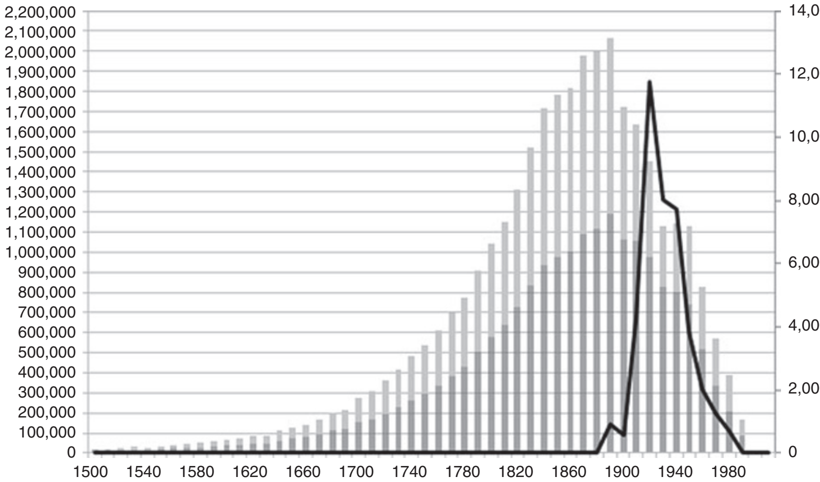

Figure 1 Frequency of birth year (dark gray) and estimated birth year (light gray) in the US genealogy (frequency on the left). Frequency of the birth year for linked veterans is shown as the black line with the frequency shown on the right.

original data sources provides insight into the quality of the linking. For example, when there is overlinking, families become too large, with too many spouses and too many children.

2. Comparison of key historical pedigrees to other sources. There are a number of well-researched and published genealogies for historical figures that are available for comparison; manual and automatic "difference" checking between pedigrees is used.
3. In addition to detail-level checking, analysis of the entire database using average coefficient of kinship, inbreeding, and relatedness mirror the numbers published for the Utah Population Database,[21] a similar Utah-based genealogy.

GenMergeDB has been used for other projects[22,23] involving genealogy creation and linking to external data. A detailed analysis of the results of a GenMergeDB record-linking project, in which a large "truth set" was available, determined that the precision of the record linking was 98–99% and the recall was 70–80%.[24] Our focus is on creating good-quality links, so the records in the genealogy are somewhat underlinked. The problems associated with using a genealogy that is missing records, or is underlinked, have been addressed in ref. 3.

VA patient data

More than 1.3 million VA patients' demographic records are available for veterans who used the VA facilities in or near Utah and Massachusetts since 1985. This VA patient population was linked to the current US genealogy data. The result of linking is a list of VA patient identifiers and their corresponding genealogy identifiers. Record linking of VA patient demographic data to the genealogy is performed behind the VA firewall by VA personnel and all VA data and identifiers remain behind this firewall. The VA patient records and genealogy records stay separate and no information from the VA records is added to the genealogy. Outside the firewall only the genealogy identifier is used. All access to health data was approved by the institutional review board as well as an oversight committee for the VA data resource.

Pedigree analysis

With a resource that combines genealogy data with patient data, it is straightforward to identify clusters (or pedigrees) including related VA patients with a specific phenotype of interest (e.g., all patients diagnosed with prostate cancer). Using well-developed and published methods, we can identify those phenotypes that are observed to occur in relatives more than expected.[25]

We can test the hypothesis of "no excess relatedness" for a phenotype of interest by estimating the average relatedness of all possible pairs of VA patients diagnosed with the phenotype and comparing it with the expected relatedness of randomly selected, matched VA patients. We can also estimate the relative risk for a specific phenotype among the close and distant relatives of cases, using phenotype rates estimated within the VA patient set. In addition, we can identify high-risk pedigrees. A high-risk pedigree is defined as the set of all descendants of an individual among whom there is a statistically significant excess of the phenotype of interest based on disease rates in the VA resource. We propose using these analyses to consider disease- (e.g., posttraumatic stress disorder) and health-related (e.g., longevity or body mass index) phenotypes included in the genealogy or medical data.

RESULTS

Genealogy

The record linking of these genealogy sources has resulted in a resource with 22.5 million linked individuals. The current genealogy includes more than 3.5 million people with an explicit birth date after 1900 and more than 6.9 million people with an estimated birth year after 1900. A large percentage of publicly available genealogy data does not include a birth year

Table 1 Total Veterans Administration (VA) patients and linked VA patients by facility and state

Massachusetts area facilities	Total patients	Linked	Percentage
(402) Togus, ME	119,242	1,809	1.5
(405) White River Junction, VT	74,687	1,088	1.5
(518) Bedford, MA	58,142	486	0.8
(523) Boston HCS (Boston)	281,801	2,473	0.9
(608) Manchester, NH	67,565	750	1.1
(631) Northampton, MA	49,474	490	1.0
(650) Providence, RI	101,128	793	0.8
(689) Connecticut HCS (Westhaven)	215,449	1,482	0.7
Total Massachusetts	967,488	9,371	1.0
Utah area facilities			
(436) Montana HCS (Fort Harrison, MT)	95,243	2,459	2.6
(442) Cheyenne, WY	63,705	1,475	2.3
(575) Grand Junction, CO	39,052	1,124	2.9
(660) Salt Lake City HCS (Salt Lake City, UT)	186,561	25,861	13.9
(666) Sheridan, WY	36,299	1,000	2.8
Total Utah/intermountain states	420,860	31,919	7.6
Overall VA total	1,388,348	41,290	3.0

Table 2 Example counts of VA patients with genealogy data, by ICD-9 coding

ICD-9 code	Disease definition	No. of linked VA patients
401.9	Essential hypertension	17,544
272.4	Hyperlipidemia	15,633
530.81	Esophageal reflux	9,785
250.00	Diabetes mellitus without complication	7,807
724.2	Lumbago	7,983
278.00	Obesity	6,141
366.16	Senile nuclear sclerosis (cataract)	5,440
780.57	Sleep apnea	4,394
389.11	Sensory hearing loss	4,291

ICD, International Classification of Disease; VA, Veterans Administration.

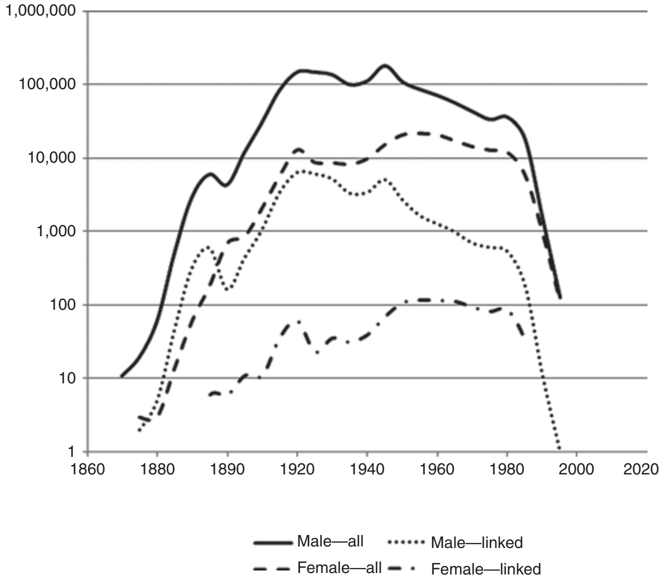

Figure 2 **Frequency of the set of 1.4 million veterans available for record linking as compared with the 41,290 veterans who linked to the genealogy by sex and birth year.**

due to lack of data for historical populations and also to censoring of birth dates for living populations. The estimated birth year is computed for those individuals without an explicit birth year, based on birth, death, and marriage dates of relatives. For example, the birth year of a parent is assumed to be 20 years before the birth year of a child. This gives a rough estimate of the birth year that helps to prevent errors in record linking caused by linking individuals who could not have lived in the same time period. **Figure 1** shows the frequency of individuals by birth year (dark gray) and estimated birth year (light gray). The genealogy population peaks at birth years near 1900 and then declines; the frequency distribution of the linked VA patients is shown in a black line and peaks at ~1920. This illustrates the lower availability of precise birth year for currently living individuals and for historical populations and the overlap of the VA and genealogy populations.

VA patients

To date, we have linked 41,290 (3%) of 1.3 million available VA patients to an individual in the US genealogy. Most of the linked patients are male (87%). Overall, we linked 3.3% of the males (40,385), but only 0.5% of the females (905). This is probably due to surname changes that affect women but not men; women are often listed in patient data with their married name but in genealogy data with their maiden name. **Table 1** lists the facilities and the total number of patients from VA facilities in Utah and Massachusetts, and shows the number and percentage linked. As expected, even though there are twice as many patients in the Massachusetts area facilities as in the Utah facilities, we linked records in Utah at a higher rate (7.6%) than in Massachusetts (1.0%); this is probably because Utah has been our primary focus for genealogy data collection to date. We observed very similar sex and birth year distributions for the 41,290 linked veterans and the 1.4 million total veteran's records available to be linked (**Figure 2**).

The fields available for record linking include name, birth date, death date, sex, and mother's name. GenMergeDB uses a single threshold score that is initially estimated by the system using the discriminating power and frequencies of the different field values. After an initial run, the links close to the cutoff are checked manually to verify that the links are valid. This cutoff score is changed to a higher value if suspect links are found.

Relationship information (for the patient's mother) is treated differently than a value for a name or date. The matching is handled by the family scoring. The VA patient record contained a mother's name two-thirds of the time. This is often

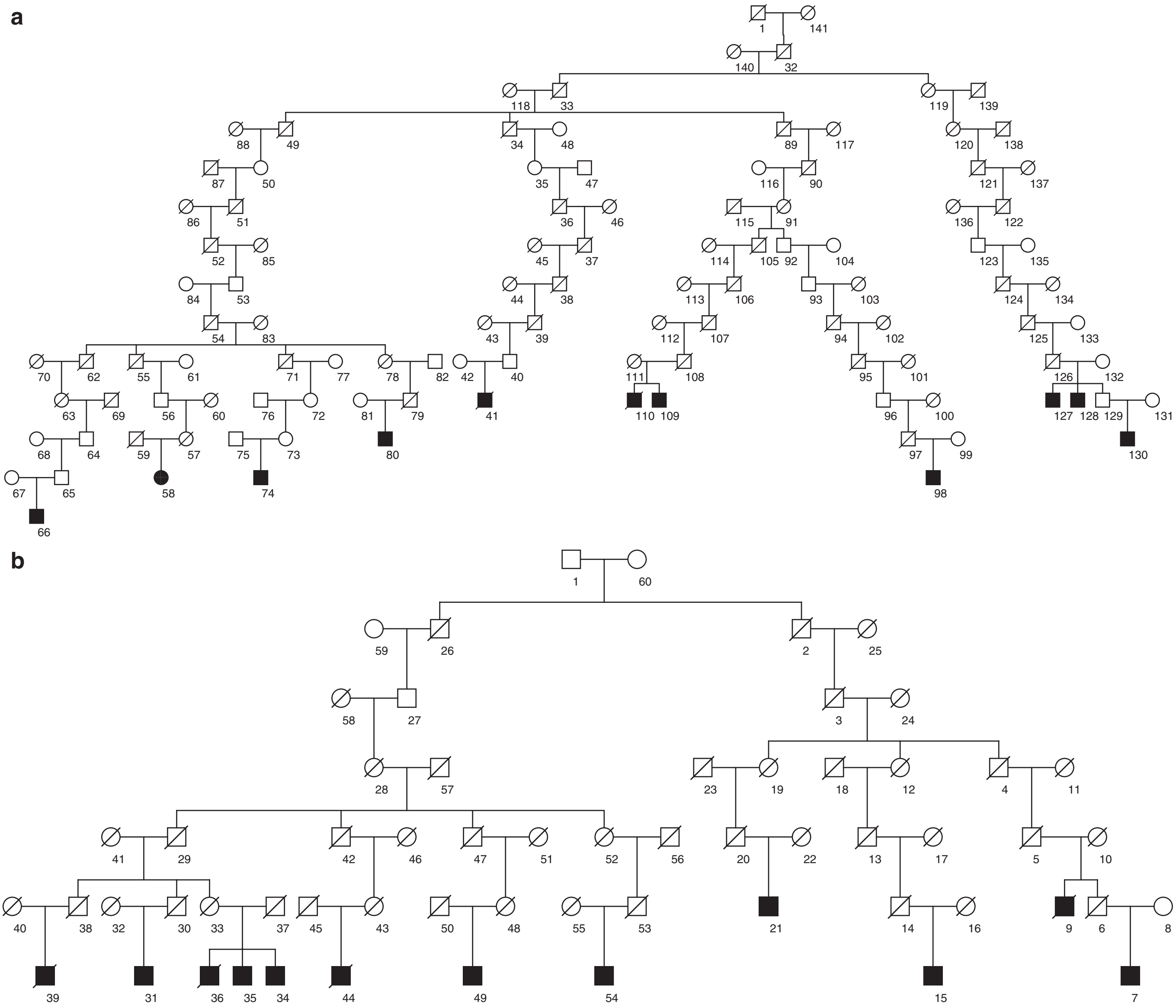

Figure 3 Example patient pedigrees. (**a**) Pedigree 1: Massachusetts Veterans Hospital Administration (VHA) patient pedigree. The founder of pedigree 1 was born in ~1500 and has 163,337 descendants in the US genealogy, including 11 Massachusetts area veterans (filled squares). (**b**) Pedigree 2: Utah VHA patient pedigree. The founder of pedigree 2 was born in England in ~1700 and has 1,295 descendants in the US genealogy, including 12 Utah area veterans (filled squares).

only the mother's maiden name (94% of the mothers); some records have the first name and maiden name. The records for mothers are considered sparsely populated as they have only the name fields and usually only the last name. Even so, this additional name is important in increasing linking scores if the mother's name matches, or decreasing linking scores, if not. In the presence of a matching mother's name, we will be able to link individuals with relatively low scores based only on their demographic information.

VA phenotype data

Only the demographic data for VA patients are used to record link patients to the US genealogy; phenotype data are not stored with demographic information. To demonstrate the range of diagnosis data available, we have initially used the International Classification of Disease (ICD) Revision 9 diagnosis codes to assign phenotypes to patients. We counted patients by diagnosis (using the presence of a single ICD-9 code for phenotype) to get an idea of representation. Among the linked VA patients, there were 273 different phenotypes (that we associated with a single ICD-9 code, e.g., diabetes = ICD-9 250.00) that were diagnosed in 1,000 or more patients. Some examples of such disease phenotypes with the largest sample sizes are shown in Table 2. Analysis of any specific phenotype could integrate other ICD-9 codes and other stored medical data (e.g., medications or laboratory tests) to more accurately identify individuals with a phenotype of interest; such phenotype assignments will be made in collaboration with the VA Informatics and Computing Infrastructure project, also located in the George E. Wahlen Department of Veterans Affairs Medical Center in Utah. VA Informatics and Computing Infrastructure is focused on providing researchers access to integrated national VA data sets as well as tools for analysis of VA data. The project utilizes data elements such as ICD and Current Procedural Terminology codes, demographics, and concepts in text documents and includes methods such as natural language processing to assist with appropriate assignment of health-related phenotypes within VA patient data.

Pedigree analysis

We do not present any analyses for specific phenotypes in this initial report. We have identified all clusters of related VA patients from the Utah and Massachusetts facilities in which all descendants of a cluster have a single common ancestor (also termed a pedigree). Thousands of such "US service" pedigrees with at least three related linked patients have been identified; 1,223 clusters of size 3; 932 clusters of size 4; 615 clusters of size 5; 1,734 clusters of size 6–10; and 6 clusters of size 200 or larger. Because the entire population of linked VA patients was used to identify these clusters, it is not possible to appropriately use the analysis techniques described to determine whether these pedigrees represent evidence for a heritable contribution (rather than representing, for example, behavioral clustering) to "service in the US Armed Forces," although such an effect is possible, if not likely. **Figure 3** shows two medium-sized example "US service" pedigrees; the pedigrees have been trimmed to reduce their size. These pedigrees represent the sort of pedigrees we will analyze to test for a genetic contribution to health-related phenotypes.

DISCUSSION

We have begun creation of a US genealogy linked to the population of US veterans receiving medical care from the Veterans Hospital Administration. Creation of the genealogy does not require any identification of VA patients and their medical data outside the VA system. We will use this combined biomedical resource to study the genetics of multiple health-related phenotypes, including phenotypes of specific significance to the VA population that cannot easily be studied in any other population (e.g., posttraumatic stress disorder).

As seen in **Figure 3**, the pedigrees identified in this resource may be many generations deep, but will only include affected individuals identified in the years for which data are available. For phenotypes identified in the VA medical data (e.g., prostate cancer), this may be limited to a few generations; for phenotypes identified in the genealogy data (e.g., longevity) many more generations will have informative data. The largely horizontal familial aggregation represented for most of the medically derived phenotypes we will analyze is the strength of this resource for genetic research. It is the distant relationships that will be informative for evidence of familial clustering (excess distant relatedness being more likely for a genetic contribution). This resource will be most valuable for studies of shared genetic effects. In potential future family studies, the distant relationships would be most informative to identify regions of chromosomal sharing.

As the genealogy increases in size and extends to other states, pedigrees will increase in both size and in the genetic distance between cases, and will be linked into larger related units, thereby increasing the power for genetic studies. Integration of other data, including geographical locations and environmental exposures, will extend the types of research that can be accomplished to epidemiologic and gene/environment studies. In contrast to the other large genealogy/medical data resource in the United States (the Utah Population Database), we hope to eventually integrate the information contained in the VA resource into clinical care. This could initially be based on, for instance, disease-risk estimates and screening recommendations derived from individual–family history for specific diseases. Although the Utah Population Database has been used to estimate family history–specific risks,[26] the protections that make genetic research with the database possible also prohibit its direct clinical use.

Analysis of familial clustering of some disease phenotypes is under way; we have limited the presentation of results to the identification of multiple extended pedigrees representing US veterans served by the VA. These "US service" pedigrees are interesting in that they represent families in which participation in one or more branches of service or one or more US wars has included members related across multiple generations in multiple different lineages. These service pedigrees are both wide and deep. Such pedigrees may represent random (expected) familial clustering, given the rates of service and participation in VA medical care in the United States. Regardless, it is clear that there are extended military families with multiple members receiving care through the Veterans Hospital Administration. In the future, this resource could be used with respect to clinical decisions related to the utility of genetic testing. With appropriate permissions in place, results of tests for one veteran might be informative in directing the care of other veterans within the VA population.

The VA genealogy/biomedical resource does have some limitations. Genealogy data quality depends on the care and attention with which it was curated (although typically the process of building a genealogy is to uncover true relationships). The genealogy data used in this resource were limited to those that are publically available; often the living members of families are censored. Relationships between individuals who are not identified by record linking are censored. Similarly, health-related phenotype data would be censored by failure to identify VA patients in the genealogy data. VA patient data only exist from 1985 to the present; all previous health data are censored, as are any phenotype data for patients seen in different health-care facilities.

Assignment of phenotypes for VA patients in this resource will only be as good as the VA medical data allow; limitations in precise assignment of phenotype from existing medical data are recognized. Parallel efforts in Utah within the VA Informatics and Computing Infrastructure program are focused on such phenotype assignment using as much data as possible[27] and will be integrated into the analysis of this resource.

The record-linking rate for VA patients using the Utah or Massachusetts VA facilities in this second year of the project was low, 3% overall, but this will increase as genealogy data are added. Even this low linking percentage has provided us with tens of thousands of VA patients with both medical data and genealogy data. Many factors contributed to the linking rate, including incomplete genealogy data, especially for the most recent birth cohorts; lack of data items in the VA patient demographic data that aid record linking; and the recent birth year constitution of much of the VA patient population. As the resource continues to grow, we will expand the genealogy data and record-linking rates will improve. There are few risk factors for diseases that would be expected to affect record-linking success; therefore, the set of linked records can be assumed to represent the VA population in an unbiased, if incomplete, fashion.

This current Veterans Genealogy Project already represents the largest resource of its type in the world and includes only two US states. With the large geographical range, extensive ethnic/racial variation represented in the VA population and in the US genealogy, and the range of environmental exposures represented in the VA population, this resource has the potential for unsurpassed opportunities for genetic studies. To enhance its utility for risk prediction, we propose that future expansions of this VA resource allow individuals to add to their own genealogy data as well as personal health history (with appropriately limited access and use).

The recent initiation of the VA Million Veterans Program, which has begun sampling DNA for all veterans using the VA medical system, will soon result in the largest genetic biorepository in the world. Combination of the VA Million Veterans Program biorepository (supplemented with recent genealogy data for the veterans sampled) with the VA genealogy/biomedical resource described here would provide incredible potential to identify and understand both common and rare genetic variants and their association with health-related phenotypes.

ACKNOWLEDGMENTS

This study was supported by an award from the Veterans Health Administration.

DISCLOSURE

The authors declare no competing interests.

REFERENCES

1. Skolnick MH. The Utah genealogical database: a resource for genetic epidemiology. In: Cairns J, Lyon JL, Skolnick M (eds). *Banbury Report No. 4: Cancer Incidence in Defined Populations*. Cold Spring Harbor Laboratories: New York, NY, 1980:285–297.
2. Cannon L, Bishop DT, Skolnick MH, Hunt S, Lyon JL, Smart C. Genetic epidemiology of prostate cancer in the Utah Mormon genealogy. *Cancer Surv* 1982;1:48–69.
3. Cannon-Albright LA, Thomas A, Goldgar DE, et al. Familiality of cancer in Utah. *Cancer Res* 1994;54:2378–2385.
4. Weires MB, Tausch B, Haug PJ, Edwards CQ, Wetter T, Cannon-Albright LA. Familiality of diabetes mellitus. *Exp Clin Endocrinol Diabetes* 2007;115:634–640.
5. 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.
6. Albright FS, Orlando P, Pavia AT, Jackson GG, Cannon Albright LA. Evidence for a heritable predisposition to death due to influenza. *J Infect Dis* 2008;197:18–24.
7. Gottfredsson M, Halldórsson BV, Jónsson S, et al. Lessons from the past: familial aggregation analysis of fatal pandemic influenza (Spanish flu) in Iceland in 1918. *Proc Natl Acad Sci USA.* 2008;105:1303–1308.
8. Arnar DO, Thorvaldsson S, Manolio TA, et al. Familial aggregation of atrial fibrillation in Iceland. *Eur Heart J* 2006;27:708–712.
9. Gudbjartsson T, Jónasdóttir TJ, Thoroddsen A, et al. A population-based familial aggregation analysis indicates genetic contribution in a majority of renal cell carcinomas. *Int J Cancer* 2002;100:476–479.
10. Sveinbjörnsdottir S, Hicks AA, Jonsson T, et al. Familial aggregation of Parkinson's disease in Iceland. *N Engl J Med* 2000;343:1765–1770.
11. Gretarsdottir S, Thorleifsson G, Reynisdottir ST, et al. The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat Genet* 2003;35:131–138. Erratum in: *Nat Genet* 2005;37:555.
12. Grant SF, Thorleifsson G, Reynisdottir I, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 2006;38:320–323.
13. Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 1994;266:66–71.
14. Kamb A, Shattuck-Eidens D, Eeles R, et al. Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nat Genet* 1994;8:23–26.
15. Tavtigian SV, Simard J, Rommens J, et al. The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nat Genet* 1996;12:333–337.
16. Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Soc* 1969;64:1183–1210.
17. Winkler WE. Overview of record linkage and current research directions, Volume 2006–2. Statistical Research Division, US Census Bureau: Washington, DC, 2006.
18. Taft RL. *Name Search Techniques.* Identification and Intelligence System: Albany, New York, New York State, 1970.
19. Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, Alexandria, VA, 1990:354–359.
20. Dintelman, SM, Maness, AT. Reconstituting the population of a Small European Town using probabilistic record linking: a case study. *Family History Technology Workshop* 2010, http://fht.byu.edu/prev_workshops/workshop10/papers/1-3-Dintelman.pdf.
21. Jorde LB. Inbreeding in the Utah Mormons: an evaluation of estimates based on pedigrees, isonymy, and migration matrices. *Ann Hum Genet* 1989;53(Pt 4):339–355.
22. Ogilvie JW, Braun J, Argyle V, Nelson L, Meade M, Ward K. The search for idiopathic scoliosis genes. *Spine* 2006;31:679–681.
23. Farrington P, Frech G, Wong L, et al. The heritability of endometriosis in a Utah population. Evolutionary and Population Genetics, Poster 585/W:245, American Association of Human Genetics Annual Meeting, Honolulu, HI, 20–24 October, 2009.
24. Schone, P. Development of an evaluation paradigm for recordmatch and its application to GenMergeDB clustering results. *Family History Technology Workshop*, 2011. http://fht.byu.edu/prev_workshops/workshop11/papers/1-1-schone%20Paper.pdf.
25. Cannon Albright LA. Utah family-based analysis: past, present and future. *Hum Hered* 2008;65:209–220.
26. Taylor DP, Burt RW, Williams MS, Haug PJ, Cannon-Albright LA. Population-based family history-specific risks for colorectal cancer: a constellation approach. *Gastroenterology* 2010;138:877–885.
27. Nelson RE, Nebeker JR, Sauer BC, LaFleur J. Factors associated with screening or treatment initiation among male United States veterans at risk for osteoporosis fracture. *Bone* 2012;50:983–988.