LIVING to 100
SOCIETY OF ACTUARIES
INTERNATIONAL SYMPOSIUM

Compiling a Very Large Sample of Centenarian Pedigrees to Ascertain Patterns of Inheritance and a "Familial Propensity for Longevity Score"

Lisa Nussbaum
Department of Biostatistics, Boston University School of Public Health
Giacomo Nebbia
Department of Medicine, Boston Medical Center and Boston University School of Medicine
Annie Helmkamp
Department of Medicine, Boston Medical Center and Boston University School of Medicine
Stacy Andersen
Department of Medicine, Boston Medical Center and Boston University School of Medicine
Thomas Perls
Department of Medicine, Boston Medical Center and Boston University School of Medicine
Paola Sebastiani
Department of Biostatistics, Boston University School of Public Health

Presented at the Living to 100 Symposium
Orlando, Fla.
January 4–6, 2017

Compiling a Very Large Sample of Centenarian Pedigrees to Ascertain Patterns of Inheritance and a "Familial Propensity for Longevity Score"

Lisa Nussbaum,[1] Giacomo Nebbia,[2] Annie Helmkamp,[3] Stacy Andersen,[4] Thomas Perls,[5] Paola Sebastiani[6]

Abstract

It is apparent that a large portion of the baby boomer population will live beyond the age of 90 years. Entitlement programs and various insurance products have thusly become interested in longevity risk. Beyond period life table predictions, actuaries have little to go on in determining which individuals or portions of populations are at increased risk of living to 90 or 100 or even older. We and others have noted strong familial risk for living beyond the oldest one percentile of survival for a birth cohort. However, just because one is at increased risk, the odds of achieving such a milestone are still small if the event is very rare. We hypothesized that determining common patterns of longevity (e.g., paternal, maternal, skipping generations) and level of risk according to which of one's relatives were long-lived can help inform actuaries about longevity risk. To explore this hypothesis, we proposed to perform network analyses of thousands of pedigrees that provide vital information for each family member. An important step of this work is to compile the largest possible samples of pedigrees with and without long-lived family members. Here, we describe our process of hand-curation of centenarian pedigrees and the software we have developed for the automated construction of such pedigrees.

Introduction

Traditionally, actuaries determine risk of lower life expectancy. Recently, though, as average life expectancies increase and as baby boomers approach retirement, entitlement programs and certain insurance products have become interested in longevity risk. Beyond sex- and race-specific cohort life tables, socioeconomic status, geographic location, years of education, absence of certain

[1] Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.
[2] New England Centenarian Study, Geriatrics Section, Department of Medicine, Boston Medical Center and Boston University School of Medicine, Boston, MA, USA.
[3] New England Centenarian Study, Geriatrics Section, Department of Medicine, Boston Medical Center and Boston University School of Medicine, Boston, MA, USA.
[4] New England Centenarian Study, Geriatrics Section, Department of Medicine, Boston Medical Center and Boston University School of Medicine, Boston, MA, USA.
[5] New England Centenarian Study, Geriatrics Section, Department of Medicine, Boston Medical Center and Boston University School of Medicine, Boston, MA, USA. Correspondence: thperls@bu.edu.
[6] Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.

diseases, and obesity and absence of unhealthy habits can help inform individual survival risk to the mid-80s. A major challenge, though, is determining who is at substantial risk of surviving up to 15–20 years beyond such ages. For an extremely smaller proportion of the population, their genetic risk is much higher (Sebastiani, Nussbaum et al. 2015). Analyses of large numbers of pedigrees containing centenarian(s) could help inform us about longevity risks of individuals who fit into specific patterns of familial longevity, and we have introduced a network-based approach to model the chance of living to extreme old age, given information about relatives' longevity (Sebastiani, Andersen et al. 2016).

The network-based model needs to be trained on a large number of pedigrees, but building accurate pedigrees is particularly laborious and often fraught with inadequate sources of information. The availability of web-based records (e.g., Familysearch.org, Ancestry.com, digitized census records, the Social Security Death Index) has revolutionized this process, but there are still major obstacles to overcome. We describe here what necessary resources are available for the manual curation of pedigrees, and how we use this data for manually validating and expanding pedigrees of centenarians. We then describe an algorithm we have implemented to automatically retrieve these data from both databases and online resources and automatically validate and expand pedigrees.

Data Sources

The New England Centenarian Study

The main source of data is the New England Centenarian Study (NECS), which began in 1994 as a biopsychosocial characterization of living to 100 with the population-based recruitment of all persons 95 and older in an eight-town area surrounding Boston, Massachusetts, and later expanded enrollment to all of North America (Sebastiani and Perls 2012). Collection of family pedigree data supported the adage that "old age runs in families": the sibling relative risk (SRR) for age of survival was elevated and increased with increasing age of the proband beyond 95 years old. Starting at an SRR of 4 with a proband living to age 95, it rose to 30 for a proband at age 106 (Sebastiani, Nussbaum et al. 2015; Perls, Wilmoth et al. 2002; Perls, Bubrick et al. 1998). To date, the NECS has enrolled approximately 2,500 centenarians and nonagenarians, including 637 semi-supercentenarians and 171 supercentenarians.

Ages of enrolled subjects are carefully validated as described in Sebastiani and Perls (2012). At the time of enrollment, we obtain information about primary family members, defined as the participant, their parents, grandparents, aunts and uncles, siblings, and children. Accurate information is often difficult to obtain and subject to inaccurate and incomplete memories. It is thus imperative to verify dates and completeness of the pedigrees.

Ancestry.com

The primary online program that we use for the manual verification/expansion of the pedigrees is Ancestry.com. Ancestry.com has searchable links to digitally available US federal and state censuses since 1790; several foreign censuses' the Social Security Death Index (SSDI); state and local

electronic birth, marriage and death registries; the Find A Grave website; war registration documents; and assorted other records that aid genealogical research. Ancestry.com provides access to approximately 12 billion US records, 2 billion available UK records and 230 million Canadian records (Ancestry.com 2015). Ancestry.com also provides many subscriber-created family trees and suggested links to supporting documents.

Vital Records

Copies of birth and death records posted on Ancestry.com are among the most important resources for construction of valid pedigrees. Birth records usually provide at least a year of birth, and they often provide exact dates of birth and sometimes the parents' names. They are generally more accurate than census data for determining the date of birth, though without parents' names, it is difficult to ascertain whether they refer to the correct individual. By providing parents' names (especially the mother's maiden and married names), the birth records are also useful in searches for brothers and sisters with the same parents.

Regarding death records, many sources are available on Ancestry.com, including the SSDI, state and local death registries, death certificates, and links to Find A Grave and obituaries. The SSDI began reporting the death of anyone with a social security number (founded in 1935) in 1936 (Hill and Rosenwaike 2002). Prior to this, there was no comprehensive US federal registry of deaths. During the early years of the SSDI, many of the deceased were not included, as few people had a social security number assigned. Further, there have been lapses in the system, and a 2001 audit detected 1.3 million deaths not included in the index. However, for years since 1973, it was estimated that between 93 and 96 percent of deaths of individuals aged 65 or older were included in the SSDI (Hill and Rosenwaike 2002). An additional obstacle emerged when a provision in the 2013 federal budget required a three-year delay between when someone has died and when their information appears in the SSDI (Felsenthal 2013).

Other Records

Ancestry.com provides links to many other types of documents, including military certificates and other records, club memberships, school records, miscellaneous public records, marriage records, city telephone directories, and immigration records, which provide additional information on family members. In particular, war registration cards and public records are useful in determining an exact date of birth, marriage records often provide the married name for women, and city directories can help estimate a date of death for a spouse.

United States Federal Census Data

The first US Census was in 1790, and censuses have taken place every 10 years since then. Because of the "72-year rule," individuals' names are available only 72 years after the census data were collected (United States Census 2016). Currently, the censuses are available from 1790 through 1940, except 1890. We accessed digitized records compiled by and provided via a generous academic agreement from FamilySearch.org.

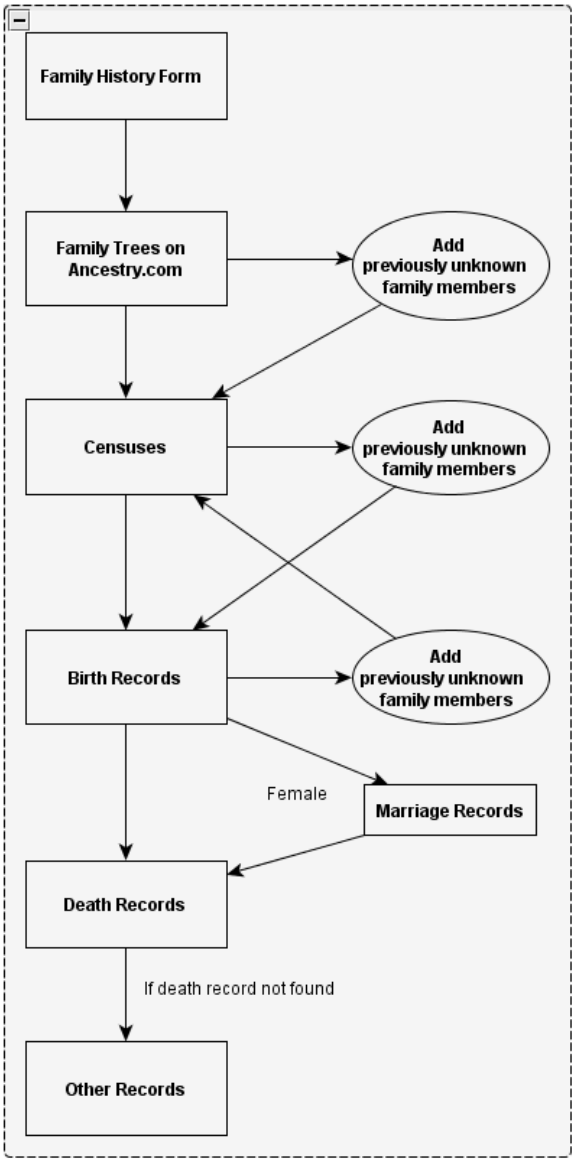
The censuses provide a great deal of information about family structure. For instance, the censuses after 1870 list the role of the member in the family: a head of the family is always reported, and all other members are labeled according to their relationship with him/her (e.g., son, daughter, wife, father). Therefore, we can ensure that the structure of the pedigree we have matches that shown in the census.

Manual Curation

Our goal in the curation process is to include all primary family members, including the probands, his siblings, their spouses, all of their offspring and spouses, parents of the proband and siblings, and the grandparents of the proband, their names, and dates of birth (DOB) and death (DOD), verified with primary source documentation whenever possible. Additionally, we try to identify the cause of death (COD).

We begin the process with information on the participant provided with the Family History Form and, if available, the Ancestry.com subscriber-created public family tree (see Fig. 1). Using information for DOB, place of birth, DOD, parents' names and spouse's name, we are often able to use one or more of these trees to add relatives we do not already have and to complete missing vital data. Trees linked to primary sources are most helpful, although we must also verify that the linked record refers to the correct person, based on available information. Information that is missing from one family tree (e.g., DOD) may be found on one of the other public family trees. For that reason, it is critical that a search be done for each member of the primary family structure who has missing or unverified data.

Fig. 1 Manual Curation Process



After searching family trees, we next look to the censuses. While many people may have the same name, whenever possible we use family members, ages and places of birth to pinpoint the correct census. Once all family members listed on the census are included in the pedigree, we look at Ancestry.com-suggested linked documents. While these documents are extremely useful, it is critical to ensure that they relate to the same individual.

After the census, we next look at birth records to help pinpoint the DOB and search for potential additional siblings. Finally, death records are searched for family members whose DOD has not yet been traced to a primary source.

To date (May 2016), we have collected 1,555 NECS pedigrees made up of 64,000 individuals. We have manually completed and verified 490 pedigrees of primary family members, and an additional

535 have been expanded and traced to census and Social Security Death Index (SSDI) data through automated pedigree reconstruction. Also, 15,300 individuals have been traced to census records, and approximately 12,000 individuals have been traced to death records in these pedigrees. Our experience with this "manual curation" is that the most challenging step is obtaining death records. State and local death registries vary greatly; some provide information to verify the identity of the deceased individual, such as parents' or spouse's names, while others include only the DOB or just the DOD of the individual. Additionally, many of the state and local registries are not currently available digitally.

Our most reliable sources of death records have been death certificates, Find A Grave, and obituaries. Several of the state death registries include a digital image of the death certificate, which provides more detailed information, including the parents' names, with the mother's maiden name, the spouse's name, DOB, DOD and sometimes COD (though even if a cause of death is noted, it can be inaccurate). Find A Grave is owned by Ancestry.com and is linked directly to cemetery records and often to other family members' records and obituaries. Obituaries often provide very valuable details about the individual's life, including their DOB, DOD, COD, parents' names and other relatives' names.

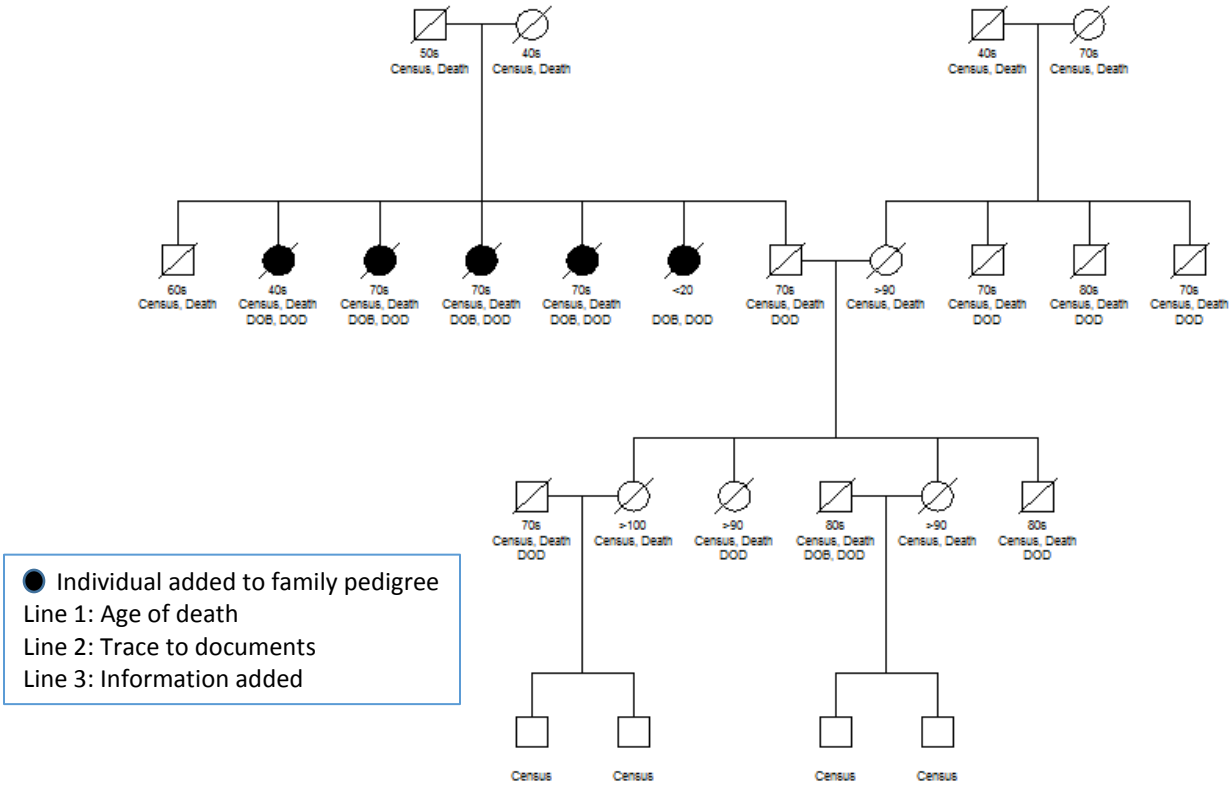
While the census does not provide a complete date of birth (except 1900, which gives the month and year of birth), it does provide the age at the time of the census, and from there, an approximate year of birth can be inferred. Since the census also provides marital status, it is useful for approximating when a spouse died. Because the census is dependent upon household reports rather than primary source documents, there is the potential for inaccuracies. We have found that the census reports closest to the birth year of the individual of interest are the most reliable for that individual. Also, if a person is recorded in a census, it can be inferred that he or she was alive at that point in time. Challenges with the census records that we have encountered include misspellings and missing family members. For these and other reasons, it is important to trace individuals carefully through multiple censuses.

Ancestry.com's subscriber-provided family trees can provide details about a pedigree not provided by censuses, such as death dates. Additionally, some pedigrees fill in data gaps caused by the absence of the destroyed 1890 US census. Family trees are also useful in finding the married and/or maiden names of female relatives. While birth records and early census data are listed under a maiden name, death records are often listed only under the married name. In the instance of multiple marriages, the family trees are often a good source of clarification of biological parentage of relatives. Moreover, it is useful to find the complete formal names of relatives, as sometimes people are listed in censuses and other documents under a blend of their first and middle names. Often these trees are created from family records, which may have more information than is currently publicly available elsewhere. However, since they are uploaded by people who are not genealogists and who are not as focused on building as complete and accurate a tree as possible, the information they contain needs to be verified with primary source documents. Many of the public

family trees are already linked to primary sources, such as census data and death records, which improve the reliability of the data presented, and when multiple trees are available, the linked ones are the ones that were most heavily relied upon.

Figure 2 shows an example pedigree. The information provided by the family showed only one paternal uncle (the individual who died in his 60s); however, five more siblings were added to the pedigree through examination of publicly available family trees and census data. There was only one individual we were not able to trace to a census or death record; the reason was that she died very young, between censuses. Three siblings in this family lived past the age of 90, and one of those lived past the age of 100. Four individuals are still alive in the children's generation.

Fig. 2 Example Pedigree



Automated Pedigree Construction

Manual curation of pedigrees requires time and diligent sleuthing. We set out to use automated computerized procedures as much as possible to search for vital data and accurately insert them into pedigrees. Giacomo Nebbia, one of the authors, developed a computer-based algorithm that builds upon initial data collected from study participants and family members. The algorithm accesses and searches several databases, including the 1850–1940 (except the 1890) US censuses (digitized and provided to us by FamilySearch.org), publicly available subscriber-assembled pedigrees and the SSDI. The subscriber-provided family trees were available through REST API,

which allows online access to the data (https://familysearch.org/developers/). The algorithm was implemented as a desktop application in Java 7, and Hibernate 3.6.10 was used to access the databases.

The algorithm first matches individuals from the NECS database to the various US census data, and then uses the set of matches in a recursive way to generate pedigrees. The search for a person in the databases is performed by requiring a perfect match on the person's sex, last name and birthplace (when available), while ± 2 years caliper matching is used for the birth year. A scoring procedure ranks the retrieved matches to select the best one. The scoring process uses techniques of "record linkage" to match records from different data sources (Fellegi and Sunter 1969). These techniques include methods for fuzzy matching, so that, for example, the name "Sarah Jane" can be matched with "Sara Jane." Such matching is enabled by a string distance function that computes a weighted average between the Jaccard index (Jaccard 1912), which considers initials only (in the example above, "S J" in both cases), and the Levenshtein distance (Levenshtein 1966), which considers full names only ("Sarah Jane" vs. "Sarah"). The implemented Levenshtein distance handles the possibility that people may have multiple names. A birth date score is also computed that allows the ± 2 years range from the searched birth date. For each retrieved census match, a final similarity value is computed as a weighted average of the name and the birth date scores.

As one can imagine, there are hundreds of thousands of people with exactly the same names and approximate birth dates, so the algorithm must effectively deal with the problem of multiple matches for the same person in the same census. To solve this issue, the algorithm uses the household members of each retrieved match and reduces the set of matches to those in which multiple household members are matched to centenarians and known relatives. This approach requires that we know at least one member of the household of a centenarian, and this is usually the case. Additional rules were implemented to allow the pedigree to grow in depth.

Once a list of confirmed individuals is assembled, the algorithm compiles a list of relationships linking pairs of individuals in the same household (Alter et al. 2009). The relations are inferred from the "relationship to head" field in the census records. The list of confirmed individuals often contained duplicates (e.g., the same individual found in different censuses). Consequently, rules to merge duplicate individuals were implemented in the algorithm, based on record-linkage techniques and on some logic rules.

After the merging phase, the algorithm generates "trios" by linking each family member to the person's parents, thus defining a pedigree. The algorithm has some procedures to flag inconsistencies in the data or incomplete information that could arise, for example, through multiple marriages. Once the final pedigree is assembled, the algorithm attempts to validate all death dates against the SSDI. Record-linkage techniques are used to rank the retrieved results.

The final step of the algorithm is to search the user-provided family trees from FamilyTree.org and use record-linkage techniques to match the reconstructed pedigrees to existing family trees. If a match is found, the family tree is used to supply additional vital details of family members (e.g.,

maiden names for women, death dates, detailed birthplaces), but no additional members are added at this stage. The final pedigree is then printed on a file in linkage format.

As an example of the potential for the developed application, starting from three documented family members (parents and child), a 42-member family tree was built using the algorithm. In each iteration, the results were manually cleaned, and the improved results were fed back to the next iteration. This process was repeated five times. Since the chosen pedigree was a very detailed one, the cleaning steps were fast, and the results could be quickly assessed.

Discussion

We have described two procedures for manual and automated curation of pedigrees. The curation includes verification of pedigree, validation and completions of DOB and DOD, and inclusion of relatives who were missing from the family history form provided by the subjects enrolled in the study. There are several caveats with the manual and automated processes detailed in this article. The biggest source of error is incorrectly matching a person in a pedigree with an entry in one of the databases. There is no fool-proof way to avoid this mistake; it is crucial to ensure that the information shown agrees or is reasonable given the other information already gathered. The use of metrics in the automated process has the advantage of limiting election bias in multiple matching.

Many of the individuals emigrated from other countries and changed their names or the spelling of their names through the immigration process. Siblings sometimes took on different spellings if they arrived at different times. Additionally, the Scandinavian method of passing on family names is different: children traditionally took the first name of their father combined with either “son” or “dotter” as their surname, which makes it more difficult to follow families through time. Obviously, there are other culture-specific naming challenges as well. Also, the censuses often have misspelled or differently spelled names, caused by the scribe writing them down incorrectly, the transcriber reading them incorrectly, or the use of different first names (for example, John and Jack) at different times. Further, for some individuals, a first name would be used in some years, a middle name during others, and initials in still other years, which makes it difficult to follow that individual through time. This gave the automated process particularly difficult issues, but each example allowed us to refine the matching process for future matches.

Neonatal deaths are often missed unless they were included in a subscriber-supplied family tree, since they often do not show up on a census. People who were born between 1880 and 1900 may be entirely missed if they died or moved away from their families before 1900, because the 1890 census is not available. In some places in the United States, there were no censuses in the early years, and families who were in transit were often not included on a census. Dates of birth and birthplace are more often incorrect and vary from one census to the next for those born outside of the United States. Another obstacle is that people born after 1940 will not appear on any available census, and there are few digitally available state and local birth registries from this time period. Moreover, some individuals were inexplicably omitted from a particular census.

The fact that a married woman will have a maiden name prior to marriage and then most often take on her husband's name after marriage poses a particularly vexing challenge. Once she has married, she is listed in the census under her married name, and her death records are also under her married name. It is therefore impossible to determine, from her name dropping off the census rolls, whether she has died or gotten married. Further, if there are multiple marriages, finding appropriate census data and death records will be even more difficult.

When no death record can be found, it is not necessarily appropriate to assume that the individual is still living. Due to the lack of a comprehensive death registry and errors in the SSDI, we were unable in numerous instances to confirm a death date. In fact, the only available death records the algorithm could search were from the SSDI, and as a result, we could confirm only 27% of the death dates. State and local death records are available for in-person viewing, but most are not currently available in a digitized online searchable form. Moreover, while these documents are available for those who were born and/or died in the United States, there are limited available records for those in other places, such as Canada or Europe, and these sources also are not currently available in a searchable format. We anticipate that as more death data such as digitized state and local records and Find A Grave become searchable online, this will become less of a problem.

For the automated search process to begin, the pedigree must contain enough initial information to ensure the algorithm can succeed. At this early stage of database development and using the algorithm, we have set this threshold at having at least one "trio" and having the birthplace, birth year, sex, surname and first name for at least 40% of the individuals in the pedigree.

Conclusions and Future Directions

It is now possible to complete and verify accurate family pedigrees for individuals in the NECS which are used for characterizing the contribution of familial longevity to an individual's chance of living to extreme ages. While the process is labor-intensive and subject to potential error, the availability of electronic data has made the process more feasible. Moreover, the development and use of computer applications to search and compile these data has the potential to greatly ease the time-intensive process.

The results of the automated process are quite promising, and the implementation of refined record-linkage techniques and merging algorithms will yield even better output. Future development also includes a graphical user interface (GUI) for a quicker and more effective analysis of the algorithm's steps, letting researchers better interact with the program. An addition that has the potential to greatly improve the overall performance is the inclusion of further data sources, especially marriage records and obituaries. As discussed, keeping track of individuals who marry can be difficult if we do not have any information about their new family; marriage records could help alleviate this problem. Additionally, such records are critical in effectively building family trees automatically.

The combination of manual and automated completion of centenarian pedigrees will provide us with a substantial database of pedigrees to conduct a variety of analyses. Multigenerational

pedigrees can be used to generate estimates of heritability of aging and extreme longevity. In addition, we have shown that multigenerational pedigrees can be used to derive a "longevity propensity score" that uses the information about relatives' longevity and additional individual specific covariates to compute the probability of an individual living to extreme old age (Sebastiani, Andersen et al. 2016). In addition to personalized prediction, such a score can be used to inform studies of human longevity and as a covariate in statistical analyses of longevity studies.

Acknowledgments

This work was supported by the National Institutes of Health–National Institute on Aging (NIA U19-AG023122), and the National Heart Lung Blood Institute (R21HL114237).

References

Alter, G., K. Mandemakers and M. P. Gutmann. 2009. Defining and Distributing Longitudinal Historical Data in a General Way Through an Intermediate Structure. *Historical Social Research [Historische Sozialforschung]* 34(3): 78–114.

Ancestry.com. 2015 (July 22). Ancestry.com LLC Reports Second Quarter 2015 Financial Results. Press release. http://corporate.ancestry.com/press/press-releases/2015/07/ancestrycom-llc-reports-second-quarter-2015-financial-results/.

Fellegi, I. P., and A. B. Sunter. 1969. A Theory for Record Linkage. *Journal of the American Statistical Association* 64(328): 1183–1210.

Felsenthal, M. 2013 (December 11). Death Master File Reform Breathes Life Into U.S. Budget Deal. Reuters. http://www.reuters.com/article/us-usa-budget-deathlist-idUSBRE9BA1D520131212.

Hill, M., and I. Rosenwaike. 2002. The Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages. *Social Security Bulletin* 64(1): 45–51, https://www.ssa.gov/policy/docs/ssb/v64n41/v64n41p45.pdf.

Jaccard, P. 1912. The Distribution of the Flora of the Alpine Zone. *New Phytologist* 11: 37–50.

Levenshtein, V. I. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics–Doklady* 10: 707–710.

Perls, T. T., E. Bubrick, C. G. Wager, J. Vijg and L. Kruglyak. 1998. Siblings of Centenarians Live Longer. *Lancet* 351(9115): 1560.

Perls, T. T., J. Wilmoth, R. Levenson, M. Drinkwater, M. Cohen, H. Bogan, E. Joyce, S. Brewster, L. Kunkel and A. Puca. 2002. Life-Long Sustained Mortality Advantage of Siblings of Centenarians. *Proceedings of the National Academy of Sciences of the United States of America* 99(12): 8442–8447.

Sebastiani, P., S. L. Andersen, A. I. McIntosh, L. Nussbaum, M. D. Stevenson, L. Pierce, S. Xia, K. Salance and T. T. Perls. 2016. Familial Risk for Exceptional Longevity. *North American Actuarial Journal* 20(1): 57–64.

Sebastiani, P., L. Nussbaum, S. L. Andersen, M. J. Black and T. T. Perls. 2016. Increasing Sibling Relative Risk of Survival to Older and Older Ages and the Importance of Precise Definitions of "Aging," "Life Span," and "Longevity." *Journals of Gerontology: Series A, Biological Sciences and Medical Sciences* 71(3): 340–346.

Sebastiani, P., and T. T. Perls. 2012. The Genetics of Extreme Longevity: Lessons From the New England Centenarian Study. *Frontiers in Genetics* 3: 277.

United States Census. 2016. The "72-Year Rule." Decennial Census Records. https://www.census.gov/history/www/genealogy/decennial_census_records/the_72_year_rule_1.html (accessed April 11, 2016).