# Digital Family History Data Mining with Neural Networks: A Pilot Study

*by Robert Hoyt, MD, FACP; Steven Linnville, PhD; Stephen Thaler, PhD; and Jeffrey Moore, PhD*

## Abstract

Following the passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, electronic health records were widely adopted by eligible physicians and hospitals in the United States. Stage 2 meaningful use menu objectives include a digital family history but no stipulation as to how that information should be used. A variety of data mining techniques now exist for these data, which include artificial neural networks (ANNs) for supervised or unsupervised machine learning.

In this pilot study, we applied an ANN-based simulation to a previously reported digital family history to mine the database for trends. A graphical user interface was created to display the input of multiple conditions in the parents and output as the likelihood of diabetes, hypertension, and coronary artery disease in male and female offspring. The results of this pilot study show promise in using ANNs to data mine digital family histories for clinical and research purposes.

## Introduction

One of the most significant scientific achievements of the past two decades was the completion of the Human Genome Project in 2003.[1] As a result, genetic links to common diseases such as age-related macular degeneration, multiple sclerosis, and Alzheimer's disease have been established.[2] Despite the treasure trove of data generated from this effort and the decreasing cost of whole-genome sequencing, multiple ethical, legal, and social challenges exist. Furthermore, because of the complexity of the human genome, significant questions remain regarding how to interpret the results. Genetic tests are best for single gene disorders with high penetrance, but they account for only a tiny percentage of chronic disorders and are therefore poor tests for screening. The reality is that most chronic diseases are polygenic disorders that have low penetrance and are influenced by multiple environmental factors. Dr. Eric Green, the director of the National Human Genome Research Institute, stated in 2011, "At the moment, the biggest challenge is in data analysis. We can generate large amounts of data very inexpensively, but that overwhelms our capacity to understand it. At the other end of the spectrum, we need to infuse genomic information into medical practice, which is really hard. There are issues around confidentiality, education, electronic medical records, how to carry genomic information throughout lifespan and make it available to physicians."[3]

While the challenges of the Human Genome Project are being addressed and clarified, some experts recommend using the routine family health history to predict future diseases and conditions. Some have referred to the family history as the "first genetic test."[4] Additionally, the information from family histories has been shown to be important for investigation of diseases with a genetic component.[5–7] For most chronic diseases, a positive family history increases the relative risk of disease in offspring two to

five times over the baseline risk, particularly if more than one first-degree relative has the condition and the age of onset is early.[8]

Prior to the adoption of electronic health records, obtaining a family history was infrequent and time consuming, and the resulting data were not structured or computable. The situation changed with the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, which established a reimbursement program for eligible professionals (EPs) and eligible hospitals (EHs) that used certified electronic health records (EHRs) and complied with meaningful use objectives.[9] As of December 2014, 509,250 EPs and 4,801 EHs had registered for the Medicare and Medicaid EHR reimbursement program.[10]

In stage 2 meaningful use, one of the menu objectives is to "record patient family health history as structured data," and the measure standard is "more than 20 percent of all unique patients seen by the EP during the EHR reporting period have a structured data entry for one or more first-degree relatives." Data standards required to support structured data are the HL7 Pedigree Standard and the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT).[11] Therefore, digital family histories are expected to emerge as part of EHRs, but what will be done with the data?

Digital family histories and whole-genome sequencing should be considered forms of clinical decision support, which is part of the EHR of the future. The goal would be to alert and inform clinicians and patients about the probabilities of future diseases and conditions. Data mining tools would be necessary to link a knowledge base with actual patient information in order to either describe a condition or make a prediction. The two main categories of data mining are supervised machine learning and unsupervised machine learning. In the former, one assumes that the data classes are known ahead of time, whereas in unsupervised learning the system is presented with data and develops classes or clusters. Supervised learning can perform predictive modeling based on dependent and independent variables, similar to logistic regression.[12]

One interesting type of data mining involves the use of artificial neural networks (ANNs) or neural networks, which are capable of both supervised and unsupervised machine learning. Neural networks use computational units that are analogous to the biological neuron. Such computational neurons are connected unidirectionally, may operate in parallel, and behave as simple switching elements, which fire when supplied a threshold level of integrated input signal. The neuron can receive multiple inputs (similar to dendrites), which are processed and weighted to generate a single output (analogous to an axon). The overall network may be viewed as a nonlinear mathematical transformation that maps input to output patterns.

In the supervised learning model, training patterns are repeatedly propagated through the net to produce outputs differing from those appearing in the training data. Such output error serves as the basis of a backward propagating wave that iteratively corrects connection weights until the net's output pattern closely matches the patterns represented in the data.[13] Neural networks are now in mainstream use, with common applications in voice and handwriting recognition. Neural networks have been applied to the field of medicine in four ways: predictive modeling, signal processing, diagnosing, and prognosticating. Neural networks have been used in almost every medical subspecialty field, such as radiology (image pattern recognition), cardiology (electrocardiogram analysis), and neurology (electroencephalogram analysis).[14]

We previously reported our experience with a digital family history collected from a unique cohort of older men who were Vietnam-era repatriated prisoners of war and a comparison group. This article builds on the previous study, published in 2013.[15] A digital family history of first-degree relatives was created using an online survey tool. The participant who took the survey reported on the health of parents, siblings, and children, and this information was exported to a spreadsheet, facilitating analysis with cross-tabulation.[16] This effort was labor intensive, so it was postulated that neural networks might be a means of mining this rich data. This pilot study reports on the results of evaluating the digital family health history database with neural networks, as compared to cross-tabulated results.

# Methods

## *Participants*

The study population consisted of 319 male Vietnam-era veterans, which included 253 who were repatriated prisoners of war as well as 66 in a comparison group, matched for gender, age, education, and combat roles in Vietnam. The average age at the time of survey completion was 70 ± 6 years. These individuals visited the Robert E. Mitchell Center for Prisoner of War Studies, located in Pensacola, Florida, on a near-annual basis. The program has been in existence since 1973, with some repatriated prisoners of war having 42 years of longitudinal physical and psychological data.[17] This project was approved by the institutional review board, and all patients signed a consent form. Of 447 potential participants who were e-mailed, 319 (71 percent) agreed to complete the survey.

The survey data included information on 2,412 individuals from three generations. With 709 children excluded, 1,703 male and female adults were included in the results. The data included 319 sets of parents (638 individuals) and 1,065 male and female offspring.

The 319 adult male survey participants reported on the health of themselves, their parents, their siblings, and their children. Figure 1 shows the breakdown of participants by generation and gender. Children were excluded because the pilot neural network was designed to include only the parents and the parents' male and female offspring.

With children excluded, the baseline prevalence of type 2 diabetes (DM) in parents and their offspring was 10 percent, that of hypertension (HTN) was 31 percent, and that of coronary artery disease (CAD) was 6 percent.

## *Survey Development*

To review the survey content and face validity, we convened an expert panel consisting of a university-based geneticist, a private genetic counselor, a neuropsychologist, and an experienced internal medicine physician to determine the appropriate survey design and the selection of common medical and psychiatric diseases with a genetic component. A literature review was also undertaken to determine the availability and relevance of existing family history questionnaires. We also benchmarked our efforts with the recommendations made by the 2008 American Health Information Community's Family Health History Multi-Stakeholder Workgroup.[18] A commercial survey instrument (SurveyMonkey) was used to create the web-based survey.[19] The survey had the following sections:

1. Demographic questions including gender, adopted status, twin status, and ethnicity, to be answered by all participants prior to proceeding. Adopted individuals were excluded.
2. Personal health information divided into the following question categories. All categories had a free-text answer option. The number of questions in each category is in parentheses. In this section only, participants used a drop-down menu to specify the age at which they received the diagnosis.
    a. General condition questions (8)
    b. Heart condition questions (5)
    c. Cancer questions (14)
    d. Brain disease/neurodegenerative disease questions (6)
    e. Mental disorder related to learning disability questions (2)
    f. Mental disorder other than related to learning disability questions (8)
    g. Substance abuse questions (2)
3. Mother's health
    a. Living/deceased (drop-down menu); current age or age of death (drop-down menu); smoker status (drop-down menu); served in military (drop-down menu).
    b. The questions from section 2 (personal health) are again asked (total of 50 questions), but there is no option to record age of diagnosis.
4. Father's health; questions identical to the mother's health section.
5. Sibling health; questions identical to the mother's health section.

6. Children's health; questions identical to the mother's health section.

The data collection period was from May 2012 to June 2013. The collection tool was online, so participants could complete the survey at home, or they could complete the survey in Pensacola, Florida (a midsized city in the Southeast region of the country), during their annual medical follow-up examination.

Further details regarding how the survey was created, tested, and privacy protected were reported in our 2013 study.[20]

### *ANN Development*

Training data were largely converted to Boolean format, with 1s and 0s respectively denoting presence or absence of a disease, whereas other data were represented as real numbers in the range between 0 and 1.

Using a proprietary neural network training package called PatternMaster, a thousand-trial ANN architecture, involving randomly generated hidden-layer architectures and learning parameters, was rapidly generated, trained, and tested on the basis of generalization accuracy using set-aside data. During this automated testing, a separate ANN learned to map all network and training parameters to the anticipated generalization accuracy. This latter net was then stochastically interrogated to determine the network architecture, learning rate, and momentum that provided the most accurate predictions.[21] This optimal net was trained to a root-mean-square prediction error of 0.01 and exported both as a spreadsheet, whose cells functioned as neurons, and as a C-code function.

The spreadsheet-based neural net allowed for transparency and rapid experimentation. The latter feature proved valuable in determining how best to vary free input parameters after certain parameters were chosen to be kept fixed in the model. To this end, we developed two approaches and wrote macros to systematically vary the free inputs. The first of these methodologies, called MonteCarlo, varied free parameters via a "loaded" computational coin flip that reflected the disease's prevalence in the training data. Therefore, to simulate a condition occurring 20 percent of the time within the data, the disease parameter was set to 1 if a random number, in the range [0, 1], fell below 0.2. The other approach, called Variational, extracted Boolean values by randomly accessing a row of the original training data and extracting values from relevant data fields.

In the end, we found that both approaches gave similar results, but the overhead of Microsoft Excel significantly slowed down the stochastic interrogation of the model to 5 to 15 seconds. To overcome the issue of execution speed, both the Excel macros and ANN C module were converted to C# and compiled into an executable having a more intuitive graphical user interface (GUI). Using the GUI, the presence or absence of a disease in either parent could be indicated by a check or a blank check box, respectively. Floating parameters could be indicated using the third state of these boxes, shown as solid blue in Figure 2.

In the pilot phase we opted to study only the offspring (sons and daughters) of the parents. A variety of chronic diseases and conditions with a genetic component could be used as input for the mother and/or father, and the risk of diabetes (DM), hypertension (HTN), and coronary artery disease (CAD) would be displayed as output for male and female offspring. Figure 2 provides an example of the output based on input consisting of common medical problems such as diabetes and hypertension.

### *Statistical Methods*

From the family history survey data, individuals were examined through cross-tabulation of a single disease type (DM, HTN, or CAD) at a time; otherwise, cross-tabulation of multiple diseases and/or any other variable resulted in too few data points for statistical analysis. Cross-tabulations were done of the effect on the offspring of mother having the disease, the father having the disease, both parents having the disease, or neither parent having the disease. Using these cross-tabulations, a series of odds ratio (OR) analyses that compared having the disease (either parent or both positive) with not having the disease (both parents negative) were then conducted with the parental effect of the disease type on the male and

female offspring. This process also included calculating the likelihood ratios, disease-type base rates, and 95 percent confidence intervals (CIs) in comparison to the ANN output for each disease type.

To elucidate further the difficulty encountered in comparing the cross-tabulation outputs because of the small number of data points with the ANN outputs, we were forced to average ANN outputs for each disease type. The ANN GUI required smoking status (never smoked, quit smoking, and current smoker) be marked in the GUI, which would lower the number of parents analyzed. So, for example, if we evaluated parents who never smoked and who were not diabetic, and compared the ANN output to the cross-tabulation, we found in the cross-tabulation only 20 parents out of a pool of 638 that fit the criteria. Those 20 parents had 20 daughters and 20 sons, so the sample size was too small for statistical analysis, making comparison with the ANN output difficult. We therefore had to average the neural network output across all smoking categories (i.e., the arithmetic mean of the 3 smoking outputs) for each of the results to maintain a sufficient number of data points in the analyses of the cross-tabulations.

## Results

Table 1 and Table 2 compare the effect of family history for three major disease types (DM, HTN, CAD) in male and female offspring, respectively.

For the male offspring, the mother having the disease significantly affected them for DM (OR = 4.11), for HTN (OR = 2.33), and for CAD (OR = 5.19). This finding means a fourfold odds increase in DM, a twofold odds increase in HTN, and a fivefold odds increase in CAD. The odds increased when both parents had the disease, particularly for HTN (OR = 5.16) and CAD (OR = 10.90). For the female offspring, although the mother having the disease significantly affected them for DM (OR = 10.43) and for CAD (OR = 4.46), just the father having the disease significantly affected them for DM (OR = 8.44) and for CAD (OR = 5.3). Having both parents with the disease significantly affected them for DM (OR = 14.78) and for HTN (OR = 7.90). No instances of both parents having CAD were found among the female offspring.

Confidence intervals were wide because of the small sample size, and likelihood ratios were small, reflecting the small sample size and the lower-than-average prevalence of chronic diseases in this unique cohort. More than half (67 percent) of the averaged neural network results fell within the 95 percent CIs of the base rates for each of the identified diseases.

We also compared the family history inheritance trends reported in the literature with our results. Odds ratios and neural networks demonstrated that the largest increase in diabetes among offspring occurred when either the mother had DM or both parents had DM. These results reflect findings reported in the literature. Although the neural network result for DM in male offspring was 0.44, the literature suggests that it may be as high as 0.50 in male and female offspring when both parents are diabetic, so it is possible that the neural network produced more accurate results.[22]

## Discussion

To our knowledge, this is the first report of data mining of a digital family history database with the use of a neural network simulation. Our model is based on training for multiple inputs, but the output was limited to only three common disease entities, chosen because of their high prevalence and widely reported genetic component. A fully operational model would include more outputs and perhaps the ability to incorporate risk factors of both the parents and offspring. The results using neural networks correlate in general with cross-tabulation results and the medical literature, but are limited by the small sample size and low prevalence of chronic disease.

The evidence thus far indicates that inclusion of the family history has several potential benefits in healthcare. The family history can identify genetic trends, even before specific gene variants or single nucleotide polymorphisms are identified. For example, a family history of chronic obstructive pulmonary disease (COPD) is a strong risk factor for the development of COPD in offspring, in the absence of any culprit genes identified thus far.[23] Also, evidence suggests that smoking increases the risk of developing type 2 diabetes in the individual[24] and fetal exposure to smoking by either parent increases the risk of

obesity and type 2 diabetes downstream in adult female offspring,[25] presumably through an epigenetic mechanism.[26] Using the neural networks model, we demonstrated a twofold increase in diabetes in male offspring if the mother or both parents smoked. Because of the small sample size, we were not able to reliably compute the same with cross-tabulation.

The family history should also assist in population and public health, particularly in assessing the future risk of cancer and common chronic diseases, which have a genetic component.[27–30] Of note, when both parents were negative for DM, HTN, or CAD the inheritance in male and female offspring was lower than the base rate, demonstrating the high specificity of the family history, which has been reported.[31] Furthermore, the family history should aid in patient education because studies have shown that patients frequently have an inaccurate idea of their future risk of cancer based on their family history.[32] Lastly, with the movement toward "personalized medicine" and "precision medicine," both genomic sequencing and data mining of the family history are likely to be helpful in tailoring medical treatments.[33] The use of family history as clinical decision support is in its infancy and to our knowledge is not available as part of any commercial EHR system. All previous research using family histories as clinical decision support has involved standalone programs, not integrated with EHRs.[34]

Limitations of the family history should be pointed out. Collecting and maintaining a family history takes time, although using patient portals to input patient histories may lessen the burden on clinicians. Family histories may be inaccurate and subject to recall bias and may be limited by a patient's low educational status or poor family communication. The National Institutes of Health held a conference in 2009 regarding the role of the family history in improving health. Among the conclusions was that the use of family histories for predicting common conditions has low sensitivity and predictive ability but high specificity (that is, it is better for ruling out conditions).[35] Additionally, evidence suggests that knowing the family history may have only a modest effect on changing behavior.[36]

The actual database we used for training the neural networks also had a limitation. The participants who took the survey were male Caucasians with a high socioeconomic status and a low prevalence of common chronic diseases. Also, there were significantly fewer female siblings than male siblings, for unknown reasons. Importantly, the database included 2,415 individuals, but when multiple filters were applied, the actual sample size available for data mining was frequently small.

Neural networks provide an interesting alternative to other prediction models such as logistic regression. Both can be utilized for dichotomous outcomes. Neural networks are not limited by a constrained mathematical relationship between the dependent and independent variables, and they can therefore model complex nonlinear relationships. Our evaluation of neural networks was limited by choosing single disease entities in the parent, such as diabetes, without other common comorbidities, which is not realistic. Neural networks also have limitations such as the requirement of significant computational resources and the potential for model "overfitting"; also, the model development tends to be empirical.[37] Moreover, in a study comparing logistic regression with ANNs, Clermont et al. noted that the sample size needed to be in the range of 1,200 for adequate prediction from either method.[38] However, evidence suggests that neural networks can be very accurate, even with small data sets, but must be calculated correctly.[39] This study used a Monte Carlo simulation method, in which thousands of additional calculations were performed to improve accuracy.[40] As noted in the results section, neural network predictions regarding the prevalence of DM in offspring with a mother or both parents having DM closely matched the results found in the medical literature. Therefore, neural networks may actually be more accurate than cross-tabulations for small data sets.

An interesting new informatics development is the HL7 standard known as FHIR (Fast Healthcare Interoperability Resources). This standard will allow sharing of clinical decision support and the creation of applications (apps) that interact with EHRs. One of the FHIR resources involves family history, so apps could be developed that mine the family history data as a form of clinical decision supported linked to the EHR by an open application programming interface.[41, 42]

## Conclusion

The preliminary data from this pilot study provide evidence that neural networks may be valuable as a means to mine data from family histories for clinical and research purposes. For this approach to be used clinically, data standards such as SNOMED-CT must be in place, along with a means to integrate data with the electronic health record. Neural network software could be hosted remotely on a server and accessed through web services. Another option would be a family history analytical application that utilizes the new FHIR standard. From a research perspective, we believe that if neural networks are applied to a very large digital family history of patients reflecting the population at large, this data mining technique may uncover genetic trends heretofore unrecognized.

In the future, clinicians will likely be able to combine family history data, genomic data, and phenotypic data from the electronic health record into a more accurate method of disease prediction and personalized medicine. Further studies are warranted on larger and more typical patient cohorts to validate the accuracy of neural networks for data mining digital family histories and to establish causal relationships to chronic disease.

## Support

Dr. Thaler was supported under a grant from the Robert E. Mitchell Foundation in Pensacola, FL.

Robert Hoyt, MD, FACP, is the director of the Health Informatics Program at the College of Science, Engineering and Health at the University of West Florida in Pensacola, FL.

Steven Linnville, PhD, is a research psychologist at the Robert E. Mitchell Center for Prisoner of War Studies in Pensacola, FL.

Stephen Thaler, PhD, is the Founder and Chief Scientist at Imagination Engines in St. Charles, MO.

Jeffrey Moore, PhD, is a neuropsychologist at the Robert E. Mitchell Center for Prisoner of War Studies in Pensacola, FL.

## Notes

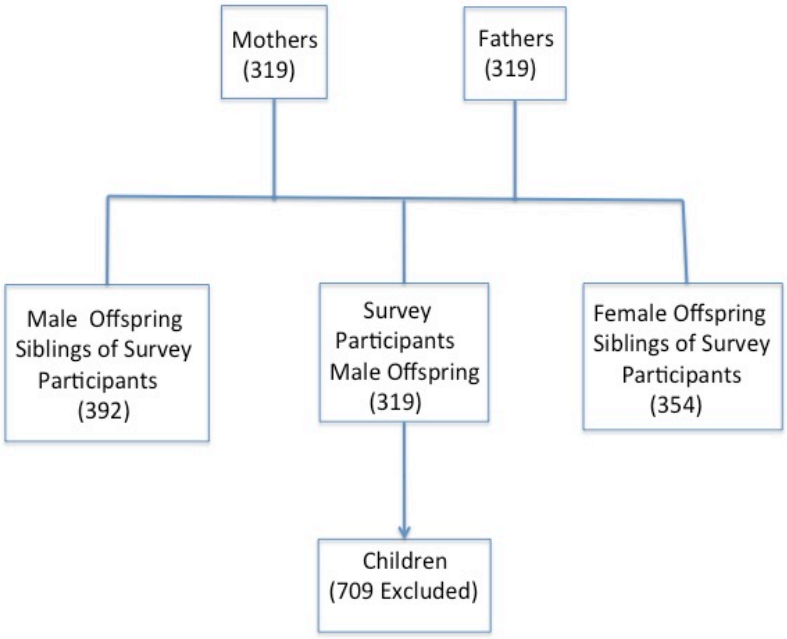

1. National Human Genome Research Institute. "All about the Human Genome Project." Available at http://www.genome.gov/10001772 (accessed February 1, 2015).
2. Bailey, Jessica N., Margaret A. Pericak-Vance, and Jonathan L. Haines. "The Impact of the Human Genome Project on Complex Disease." *Genes* 5, no. 3 (2014): 518–35.
3. Singer, Emily. "The Future of the Human Genome." *MIT Technology Review*. February 10, 2011. Available at http://www.technologyreview.com/news/422670/the-future-of-the-human-genome/ (accessed February 25, 2015).
4. Tarini, Beth A., and Joseph D. McInerney. "Family History in Primary Care Pediatrics." *Pediatrics* 132, supp. 3 (2013): S203–S2010.
5. Rich, Eugene C., Wylie Burke, Caryl J. Heaton, Susanne Haga, Linda Pinsky, Pricilla M. Short, and Louise Acheson. "Reconsidering the Family History in Primary Care." *Journal of General Internal Medicine* 19 (2004): 273–80.
6. Guttmacher, Alan E., Francis S. Collins, and Richard H. Carmona. "The Family History—More Important Than Ever." *New England Journal of Medicine* 351 (2004): 2333–2336.
7. Qureshi, Nadeem, Sara Armstrong, Paula Dhiman, Paula Saukko, Joan Middlemass, Phillip Evans, and Joe Kai. "Effect of Adding Systematic Family History Enquiry to Cardiovascular Disease Risk Assessment in Primary Care: A Matched-Pair, Cluster Randomized Trial." *Annals of Internal Medicine* 156, no. 4 (2012): 253–62.
8. Yoon, Paula W., Maren T. Scheuner, Kris L. Peterson-Oehlke, Marta Gwinn, Andrew Faucett, and Muin J. Khoury. "Can Family History Be Used as a Tool for Public Health and Preventive Medicine?" *Genetics in Medicine* 4 (2002): 304–10.
9. Title XIII. Health Information Technology. Public Law 111-5. February 17, 2009 http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/hitechact.pdf (Taccessed February 25, 2015).
10. Centers for Medicare and Medicaid Services. "Active Registrations." December 2014. Available at http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/December2014_SummaryReport.pdf (accessed February 25, 2015).
11. Centers for Medicare and Medicaid Services. "Stage 2 Eligible Professional Meaningful Use Menu Set Measures." October 2012. Available at http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/Stage2_EPMenu_4_FamilyHealthHistory.pdf (accessed February 25, 2015).
12. Hardin, J. Michael, and David C. Chhieng. "Data Mining and Clinical Decision Support Systems." In Etta S. Berner (Editor), *Clinical Decision Support Systems Theory and Practice*. New York: Springer Verlag, 2007, pp. 44–63.
13. Centers for Medicare and Medicaid Services. "Stage 2 Eligible Professional Meaningful Use Menu Set Measures."
14. Kumar, Senthil S., and Ananda K. Kumar. "Neural Networks in Medical and Healthcare." *International Journal of Innovation and Research Development* 2, no. 8 (2013): 241–44.
15. Hoyt, Robert E., Steven Linnville, Hui-Min Chung, Brent Hutfless, and Courtney Rice. "Digital Family Histories for Data Mining." *Perspectives in Health Information Management* 10 (Fall 2013).
16. Ibid.
17. Robert E. Mitchell Center for Prisoner of War Studies. Available at http://www.med.navy.mil/sites/nmotc/rpow/Pages/default.aspx (accessed February 8, 2015).
18. Feero, W. Gregory, Mary B. Bigley, and Kristen M. Brinner. "New Standards and Enhanced Utility for Family History Information in the Electronic Health Record: An Update from the

American Health Information Community's Family Health History Multi-Stakeholder Workgroup." *Journal of the American Medical Informatics Association* 15 (2008): 723–28.
19. Survey Monkey. Available at http://www.surveymonkey.com (accessed February 25, 2015).
20. Hoyt, Robert E., Steven Linnville, Hui-Min Chung, Brent Hutfless, and Courtney Rice. "Digital Family Histories for Data Mining."
21 Device for the Autonomous Generation of Useful Information. US Patent 5,659,666, issued August 19, 1997.
22. American Diabetes Association. "Genetics of Diabetes." Available at http://www.diabetes.org/diabetes-basics/genetics-of-diabetes.html (accessed February 5, 2015)/
23. Hersh, Craig P., John E. Hokanson, David A. Lynch, George R. Washko, Barry J. Make, James D. Crapo, and Edwin K. Silverman. "Family History Is a Risk Factor for COPD." *Chest* 140 (2011): 343–50.
24. Spijkerman, Annemieke M., Daphne L. van der A, Peter M. Nilsson, et al. "Smoking and Long-term Risk of Type 2 Diabetes: The EPIC-Interact Study in European Populations." *Diabetes Care* 37, no. 12 (2014): 3164–71.
25. La Merrill, Michele A., Piera M. Cirillo, Michilou Krigbaum, et al. "The Impact of Prenatal Parental Tobacco Smoking on Risk of Diabetes Mellitus in Middle-aged Women." *Journal of Developmental Origins of Health and Disease* 6, no. 3 (2015): 242–49.
26. Burgio, Ernesto, Angela Lopomo, and Lucia Migliore. "Obesity and Diabetes: From Genetics to Epigenetics." *Molecular Biology Reports* 42, no. 4 (2015): 799–818.
27. Yoon, Paula W., Maren T. Scheuner, Kris L. Peterson-Oehlke, Marta Gwinn, Andrew Faucett, and Muin J. Khoury. "Can Family History Be Used as a Tool for Public Health and Preventive Medicine?"
28. Valdez, Rodolfo, Paula W. Yoon, Nadeem Qureshi, Ridgely F. Green, and Muin J. Khoury. "Family History in Public Health Practice: A Genomic Tool for Disease Prevention and Health Promotion." *Annual Review of Public Health* 31 (2010): 69–87.
29. Claassen, Liesbeth, Lidewij Henneman, A. Cecile J. W. Janssens, Miranda Wijdenes-Pijl, Nadeem Qureshi, Fiona M. Walter, Paula W. Yoon, and Danielle R. M. Timmermans. "Using Family History Information to Promote Healthy Lifestyles and Prevent Diseases; A Discussion of the Evidence." *BMC Public Health* 10 (2010): 248.
30. Eberl, M. M., A. Y. Sunga, C. D. Farrell, and M. C. Mahoney. "Patients with a Family History of Cancer: Identification and Management." *Journal of the American Board of Family Practice* 18 (2005): 211–17.
31. Berg, Alfred O., Macaran A. Baird, Jeffrey R. Botkin, et al. "National Institutes of Health State-of-the-Science Conference Statement: Family and Improving Health." *Annals of Internal Medicine* 151 (2009): 872–77.
32. Audrain-McGovern, Janet, Chanita C. Hughes, and Freda Patterson. "Effecting Behavior Change: Awareness of Family History." *American Journal of Preventive Medicine* 24 (2003): 183–89.
33. Welch, Brandon M., and Kensaku Kawamoto. "Clinical Decision Support for Genetically Guided Personalized Medicine: A Systematic Review." *Journal of the American Medical Informatics Association* 20, no. 2 (2013): 388–400.
34. Welch, Brandon M., Kensaku Kawamoto, Brian Drohan, and Kevin S. Hughes. "Clinical Decision Support for Personalized Medicine." In Robert A. Greenes (Editor), *Clinical Decision Support: The Road to Broad Adoption*. 2nd ed. London, England: Elsevier, 2014, 383–413.
35. Berg, Alfred O., Macaran A. Baird, Jeffrey R. Botkin, et al. "National Institutes of Health State-of-the-Science Conference Statement: Family and Improving Health."
36. Ruffin, Mack T., Donald E. Nease, Ananda Sen, Wilson D. Pace, Catherine Wang, Louise S. Acheson, Wendy W. Rubinstein, Suzanne O'Neill, and Robert Gramling. "Effect of Preventive Messages Tailored to Family History on Health Behaviors: The Family Healthware Impact Trial." *Annals of Family Medicine* 9 (2011): 3–11.
37. Tu, Jack V. "Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes." *Journal of Clinical Epidemiology* 49, no. 11 (1996): 1225–31.

38. Clermont, Giles, Derek Angus, Stephen Dirusso, et al. “Predicting Hospital Mortality for Patients in the Intensive Care Unit: A Comparison of Artificial Neural Networks with Logistic Regression Models.” *Critical Care Medicine* 29, no. 2 (2001): 291–96.
39. Lanouette, Robert, Jules Thibault, and Jacques Valade. “Process Modeling with Neural Networks Using Small Experimental Datasets.” *Computers & Chemical Engineering* 23, no. 9 (1999): 1167–76.
40. Muthoni, Gateri, Stephen Kimani, and Joseph Wafula. “Review of Predicting Number of Patients in the Queue in the Hospital Using Monte Carlo Simulation.” *International Journal of Computer Science* 11, no. 2 (2014): 219–26.
41. Fast Healthcare Interoperability Resources (FHIR). Health Level 7. Available at http://www.hl7.org/implement/standards/fhir/ (accessed February 27, 2015).
42. SMART Platforms. Available at http://smartplatforms.org/smart-on-fhir/ (accessed February 27, 2015).

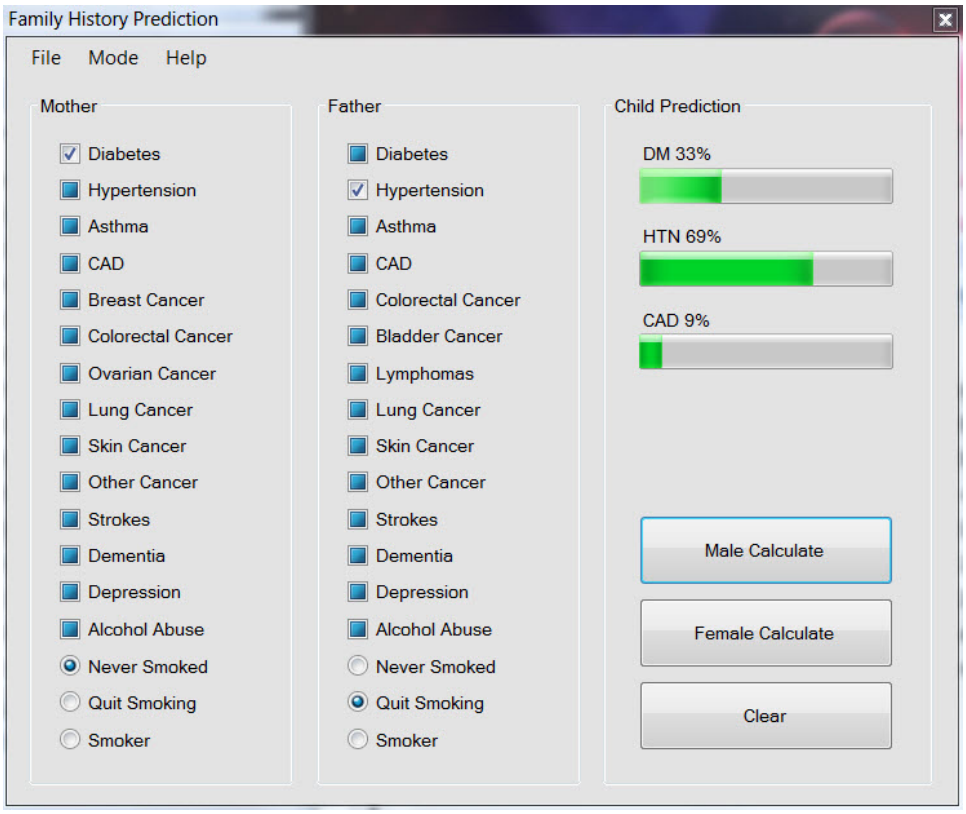
## Figure 1

Study Participants by Generation and Gender



## Figure 2

Prediction Model for Male Offspring of Mother Who Has Type 2 Diabetes and Never Smoked and Father with Hypertension Who Quit Smoking




*Note:* Floating input parameters are indicated with solid blue boxes. Output is represented by the green sliding bars on the right.

## Table 1

Comparison of Cross-tabulations with Neural Networks in Male Offspring (*n* = 711)

| | Likelihood Ratio | Odds Ratio | OR 95% CI (low, high) | Base Rate | BR 95% CI (low, high) | Neural Network |
|---|---|---|---|---|---|---|
| **DM** | | | | 0.10 | 0.02, 0.18 | |
| Any famhx | 2.43 | 3.55 | 2.14, 5.88 | --- | --- | --- |
| Specific hx | | | | | | |
| M+, F+ | 3.04 | 3.19 | 0.87, 11.75 | 0.20 | 0.00, 0.40 | 0.44 |
| M+, F- | 3.12 | 4.11 | 2.24, 7.53 | 0.24 | 0.14, 0.35 | 0.32 |
| M-, F+ | 2.53 | 2.9 | 1.36, 6.18 | 0.19 | 0.01, 0.30 | 0.22 |
| M-, F- | --- | --- | --- | 0.07 | 0.05, 0.10 | 0.12 |
| **HTN** | | | | 0.37 | 0.25, 0.50 | |
| Any famhx | 1.57 | 2.61 | 1.91, 3.57 | --- | --- | --- |
| Specific hx | | | | | | |
| M+, F+ | 3.59 | 5.16 | 3.14, 8.49 | 0.65 | 0.54, 0.75 | 0.65 |
| M+, F- | 1.77 | 2.33 | 1.56, 3.47 | 0.45 | 0.37, 0.54 | 0.46 |
| M-, F+ | 1.55 | 1.81 | 1.16, 2.81 | 0.39 | 0.30, 0.48 | 0.49 |
| M-, F- | --- | --- | --- | 0.26 | 0.22, 0.31 | 0.34 |
| **CAD** | | | | 0.11 | 0.02, 0.19 | |
| Any famhx | 2.33 | 3.00 | 1.76, 5.08 | --- | --- | --- |
| Specific hx | | | | | | |
| M+, F+ | 10.17 | 10.90 | 2.65, 44.91 | 0.50 | 0.09, 0.91 | 0.20 |
| M+, F- | 4.49 | 5.19 | 2.32, 11.63 | 0.32 | 0.14, 0.50 | 0.22 |
| M-, F+ | 1.67 | 1.82 | 0.90, 3.66 | 0.14 | 0.06, 0.23 | 0.28 |
| M-, F- | --- | --- | --- | 0.08 | 0.06, 0.11 | 0.22 |

*Abbreviations:* OR, odds ratio; BR, base rate; CI, confidence interval; DM, diabetes mellitus; HTN, hypertension; CAD, coronary artery disease; M, mother; F, father; +, positive for the disease type; -, negative for the disease type; famhx, family history; hx, history.

## Table 2

Comparison of Cross-tabulations with Neural Networks in Female Offspring (*n* = 354)

| | Likelihood Ratio | Odds Ratio | OR 95% CI (low, high) | Base Rate | BR 95% CI (low, high) | Neural Network |
|---|---|---|---|---|---|---|
| **DM** | | | | 0.08 | 0.05, 0.11 | |
| Any famhx | 3.80 | 10.12 | 4.34, 23.1 | --- | --- | --- |
| Specific hx | | | | | | |
| M+, F+ | 12.27 | 14.78 | 2.39, 91.46 | 0.33 | 0.00, 0.79 | 0.31 |
| M+, F- | 5.04 | 10.43 | 4.10, 26.57 | 0.26 | 0.12, 0.40 | 0.34 |
| M-, F+ | 5.47 | 8.44 | 2.74, 26.00 | 0.22 | 0.05, 0.40 | 0.14 |
| M-, F- | --- | --- | --- | 0.03 | 0.01, 0.06 | 0.09 |
| **HTN** | | | | 0.12 | 0.08, 0.15 | |
| Any famhx | 1.69 | 4.32 | 2.25, 8.30 | --- | --- | --- |
| Specific hx | | | | | | |
| M+, F+ | 3.89 | 7.90 | 3.47, 17.97 | 0.40 | 0.25, 0.55 | 0.37 |
| M+, F- | 2.07 | 3.55 | 1.64, 7.70 | 0.23 | 0.13, 0.33 | 0.22 |
| M-, F+ | 2.12 | 3.32 | 1.46, 7.53 | 0.22 | 0.11, 0.33 | 0.32 |
| M-, F- | --- | --- | --- | 0.03 | 0.03, 0.12 | 0.17 |
| **CAD** | | | | 0.03 | 0.01, 0.05 | |
| Any famhx | 3.18 | 5.00 | 1.47, 17.02 | --- | --- | --- |
| Specific hx | | | | | | |
| M+, F+ | 0.00 | 0.00 | --- | 0.00 | --- | 0.06 |
| M+, F- | 3.96 | 4.46 | 0.49, 40.24 | 0.08 | 0.00, 0.28 | 0.08 |
| M-, F+ | 3.58 | 5.3 | 1.43, 19.64 | 0.10 | 0.00, 0.20 | 0.15 |
| M-, F- | --- | --- | --- | 0.02 | 0.00, 0.04 | 0.06 |

*Abbreviations:* OR, odds ratio; BR, base rate; CI, confidence interval; DM, diabetes mellitus; HTN, hypertension; CAD, coronary artery disease; M, mother; F, father; +, positive for the disease type; -, negative for the disease type; famhx, family history; hx, history.