

Generation of Pedigree Diagrams for Web Display Using Scalable Vector Graphics from a Clinical Trials Database

Sanjay K. Fernando, MD, MS, Cynthia Brandt, MD, MPH, Prakash Nadkarni, MD
Center for Medical Informatics
Yale University Medical School, New Haven, CT

ABSTRACT

The standard method of studying inherited disease is to observe its pattern of distribution in families, that is, its pattern in a pedigree. For clinical studies focused on inherited disease, a pedigree diagram is a valuable visual tool for the display of inheritance patterns. We describe the creation of a web-based pedigree display module for Trial/DB, a Web accessible database developed at the Yale Center for Medical Informatics (YCFI) to support clinical research studies. The pedigree diagram is generated dynamically from the database. The icons representing each subject in the pedigree are selectable hyperlinks that will display detailed clinical data collected on the subject. Microsoft Active Server Page and Scalable Vector Graphics (SVG) are used to create the interactive pedigree diagrams.

BACKGROUND

Gregor Mendel's experiments with plants in 1865 led to the development of genetics as an experimental science. However, it was Sir Archibald Garrod who first applied the Mendelian principles of genetics to human disease in his studies of inborn errors of metabolism in 1908¹. Since then the standard method of studying inherited disease is to observe its pattern of distribution in a pedigree. The construction of a pedigree begins with the proband, the individual first evaluated in the pedigree, and extensive information about the kindred is then acquired. A pedigree diagram is a graphic that follows standard conventions to succinctly convey the distribution of disease within the members of a family.

The increasing awareness by medical researchers and physicians of the genetic contribution to disease has created a wide audience interested in studying pedigrees. Several software programs are utilized for drawing diagrams from pedigree data. The higher-end commercial packages such as Progeny² and Cyrillic³ also support management of pedigree data. These packages work very well in stand-alone mode or over a local area network but they have not been designed for sharing information over the Internet. A recently developed and freely available program,

CoPE⁴, now provides pedigree diagrams for Web display. However, all these programs (even the few that use relational database engines) support only the flat-file model of patient data. This model assumes a single row of data per patient containing numerous columns representing parameters. It is not suited for studies where the same parameters are evaluated repeatedly in a patient over time; here, a true relational model that supports multiple tables of data is more appropriate.

Clinical studies investigating the genetic basis of disease tend to increasingly require large numbers of subjects in order to improve the power of the study to detect genetic effects, and are often carried out by collaborations between geographically separated investigators. Web-accessible software for managing clinical studies data has facilitated the growth of such research. Commercial clinical studies data management systems (CSDMS) such as ClinTrial⁵ or Oracle Clinical⁶ unfortunately do not support pedigree display. Furthermore, it is not enough to simply display a pedigree diagram. The icons representing individuals should be active, in that clicking on an icon should bring up that individual's data. There is, therefore, an important need to bridge the gap between stand-alone pedigree management programs and CSDMS. This paper describes the development of an interactive pedigree display module to Trial/DB^{7,8}, an existing Web-based CSDMS developed at the Yale Center for Medical Informatics. Trial/DB is currently used to support several ongoing studies at Yale, the Vanderbilt University Cancer Center, and the National Cancer Institute supported Cancer Genetics Network (CGN), which comprises twelve medical centers nationwide.

SYSTEM DESIGN OBJECTIVES

As stated above, the pedigree display module is designed to add some of the functionality of pedigree management software to a CSDMS. The question, therefore, is how much functionality of the former must be reinvented. Currently, the CGN's collaborating centers use a variety of stand-alone pedigree management programs, so most data is bulk-imported from such programs rather than hand-entered into Trial/DB de novo. The system, however,

needs to couple the newly generated clinical/epidemiological data on the subjects in the study with their pedigree structure. Our collaborators emphasized the need to be able to navigate between the clinical data of related individuals. CSDMs typically require the user to specify a “current subject” (by typing the subject’s ID or selecting it from a list) before data on that subject can be shown. It was desired that navigation between subjects be enabled through the pedigree diagram, so that one could select the current subject through mouse clicks on appropriate icons.

A practical issue with regards to subject selection is that a large pedigree typically goes back several generations. While the ancestors, who have long been deceased, are added to the pedigree for completeness (and to increase the power of certain analyses), there is no clinical or laboratory data on them. Therefore, unlike a traditional CSDMS, it is important for the system to differentiate between “enrolled patients/subjects” who have in-depth data and “persons/non-enrolled subjects” who have only limited data. (This differentiation is also important in other circumstances where one screens a large number of individuals using a survey, before identifying a subset that matches eligibility criteria for a particular study.)

Another issue is a customizable display. When a single disease trait is being studied, there is a standard convention for displaying affected, unaffected or “carrier” subjects. (The latter have the gene abnormality but no clinical manifestations.) When one is studying the coexistence of several traits (e.g., breast cancer co-occurring with ovarian cancer) and numerous parameters are being recorded, we must allow the study designer, who is a clinician or data administrator and not a programmer, to specify which parameters to display in the diagram. (Obviously, the choice of parameters is study-specific.) Recently developed programs such as Progeny allow display of up to 4 traits simultaneously in separate quadrants of the icon representing a person: females are conventionally designated as circles and males as squares. In our own system, parameters that are binary or based on a small list of discrete values are transformed for visual display based on a color-coding scheme that the designer specifies.

INFORMATICS ISSUES

Representing a pedigree graphically poses certain interesting challenges. The alternative term for pedigree, “family tree” implies a directed acyclic graph. In practice, the pedigree may have cycles or

“loops”. This happens with a marriage between individuals who are already related through marriage or otherwise, for example, a woman marrying her uncle, or two brothers marrying two sisters. The presence of loops interferes with several mathematical analyses, as well as with the pedigree drawing algorithm, as discussed below.

The pedigree diagram is a two-dimensional representation where the X and Y coordinates of icons representing individuals must be computed based on two constraints: 1) individuals of the same generation must be at the same Y-level, with ancestors higher than descendants 2) closely related individuals must be close together on the diagram. In the scenario of uncle-niece marriage, there is an ambiguity with respect to the Y-coordinates of either individual (the woman is her own aunt). In the more common situation where remotely related individuals from widely separated parts of the pedigree marry, determining the X-coordinates of the partners is problematic. The standard method of dealing with cycles in a pedigree is to break them by “duplicating” one of the individuals participating in the cycle. That is, we have two icons for the same individual in different parts of the diagram.

IMPLEMENTATION DETAILS

Scalable Vector Graphics (SVG)⁹ is used for displaying pedigree diagrams to address the system design objectives outlined above. SVG was created by the World Wide Web Consortium (W3C), the open-standards consortium that also created hypertext markup language (HTML) and extensible markup language (XML). SVG represents an enhancement of a previous W3C-standard, vector markup language (VML). SVG is an application of XML specialized for describing two-dimensional graphics and animations.

The major benefits of using scalable vector instructions over bitmaps are the greatly reduced file size and transmission bandwidth requirement for geometric shapes, as well as resolution-independent scalability. Further, unlike bitmap images, text and images within the graphic may have its own specific hyperlink. Another benefit of SVG is that it is not proprietary like other available vector graphic formats.

Trial/DB and its Pedigree Module are implemented on a Windows NT platform. Microsoft Internet Information Server is used as the application/Web server and Oracle is the database engine (though our code can use any back-end SQL database). We use the Microsoft Active Server Pages (ASP) framework

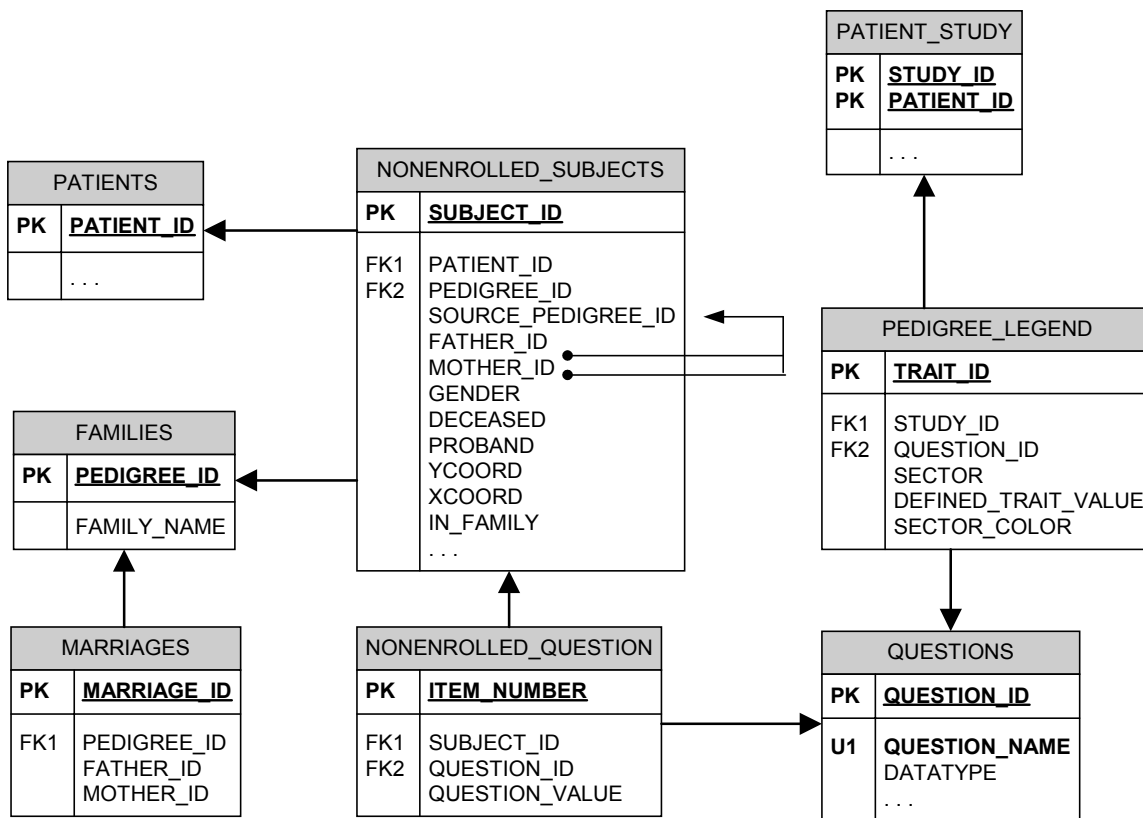


Figure 1: Database subschema for the pedigree module to Trial/DB.

for Web programming. The Web pages generated by our code also contain client-side code in the form of embedded VBScript. The server code generates SVG that is sent to the client's web browser, which requires Adobe's freely downloadable SVG plug-in for Internet Explorer and Netscape to render the graphic on the client's web browser. Figure 1 shows the subschema of Trial/DB that deals with pedigree data. (Some tables related to the rest of the system have also been shown skeletally.)

In a clinical trials database, patients are followed in a prospective fashion. Their demographic information is stored in the PATIENTS table. (Relationships between the patient's table and other clinical information in the CSDMS are not shown.) Since pedigree data contains information of deceased or non-enrolled subjects in addition to enrolled subjects, a NONENROLLED_SUBJECTS table holds pedigree data, with each family assigned a unique PEDIGREE_ID. Each member of a pedigree is assigned a sequentially generated SUBJECT_ID. However, the ID assigned to this subject in the source system from which we imported the data is also preserved, and stored in SOURCE_PEDIGREE_ID.

Every member in the pedigree data set within a family has an associated FATHER_ID and MOTHER_ID. All unique mother and father pairs in the NONENROLLED_SUBJECTS table for a particular family are then stored in the MARRIAGES table.

Consanguineous marriages and inter-lineage marriages cause complications in pedigree analysis and drawing, as previously described. To address this problem, we identify the loops in the pedigree and duplicate individuals that are involved in loops. The duplicated individual is assigned the letter "D" after their identification number. In order to determine loops, the number of lineages must first be determined in a pedigree. This is accomplished by using the IN_FAMILY field in the NONENROLLED SUBJECTS table that is initially set to zero for the family under consideration. The table is then queried to find all terminal ancestors. Each terminal ancestor represents one lineage and is successively incremented in the IN_FAMILY field. The table is then recursively queried to find all children of the terminal ancestors and each line is tagged with their ancestors IN_FAMILY value. All spouses of tagged

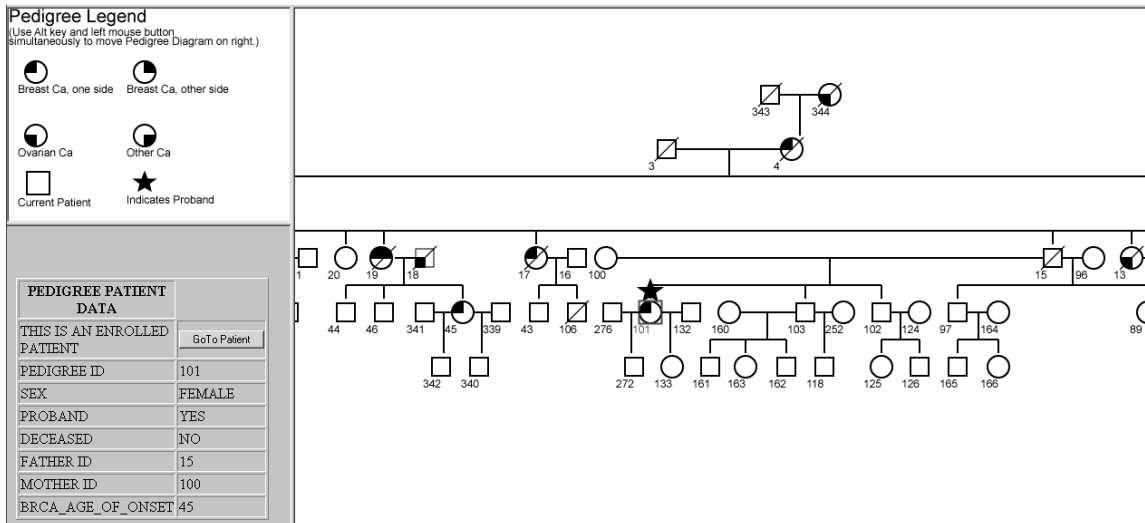


Figure 2: Pedigree diagram generated using SVG.

individuals should have an IN_FAMILY value of zero. If any spouse of a tagged individual has a non-zero tag, then a consanguineous marriage has been identified. The investigator then duplicates one of the individuals in the marriage. For all non-looped marriages the IN_FAMILY value is set to a pre-defined value to indicate that the person was married into the family.

The next loops to search for are non-consanguineous loops or inter-lineage loops. For example, a father marrying his son's mother-in-law would represent an inter-lineage loop. These marriages are obtained by querying the MARRIAGES table and comparing all IN_FAMILY values. If individuals in a lineage are linked to more than two lineages, then one of the individuals will need to be duplicated by the investigator.

The icons for all individuals on the pedigree diagram have been divided into four sectors as displayed in Figure 2. Each sector is associated with a parameter/question. For example, if a family member had breast cancer on the right side, then the individual would be shaded in the top-right sector. The definitions of the parameters of interest are stored in the QUESTIONS table of the CSDMS. The results/responses of these questions are stored in the NONENROLLED_QUESTION DATA table. The investigator or study administrator does the mapping of questions to sectors. These mappings are stored in the PEDIGREE_LEGEND table, and are study-specific. For example, a patient enrolled in a breast cancer study would have certain traits that are of particular interest to the investigator. However, if the same subject was also enrolled in another study for

colon cancer then the investigators of that study might have different traits that are of interest to them. Selection of the appropriate study would then display the pedigree diagram with the related traits.

SYSTEM FEATURES

Figure 2 is a monochrome representation of a pedigree diagram of a family consisting of over 300 relatives. (As stated previously, it is actually rendered in colors specified by the investigator. The program automatically generates the pedigree legend on the left top corner, based on this specification.) The layout is designed to minimize crossover of lines to decrease confusion. The algorithm allows for multiple marriages and intermarriages both within and between generations. To provide investigators with an easy way to identify all individuals in the pedigree diagram that are also enrolled as patients in Trial/DB, the pedigree identifications of those individuals have been highlighted in red.

Viewing clinical data on a person is as simple as selecting the person's symbol. The data on that individual is then displayed in the left lower corner. In Figure 2, the person with PEDIGREE_ID 101 is also enrolled as a patient in Trial/DB. The investigator may then select the GoTo Patient button to obtain clinical study data on the patient. Through SVG, the interactive pedigree diagram provides a rapid and comprehensive view of the clinical trials data that is not possible with a printed copy of the pedigree diagram.

Investigators can view family relationships and manage patient data from the pedigree diagram, and perform cross-pedigree searching and reporting

within a study. The pedigree diagrams and data tables can be displayed simultaneously. The relationship between the diagram and tables are dynamically linked so that changes made to the data tables will be automatically reflected in the diagram.

FUTURE WORK

Our goals for future development involve the development of an export module of the pedigree data for linkage analysis in collaborative software packages. We are also planning on developing a user interface that would allow those with minimal computer experience to select traits and import pedigree data from collaborative software packages.

ACKNOWLEDGMENTS

The pedigree data used for constructing the pedigree diagram in Figure 2 is based on the Breast/Ovarian cancer pedigree studies done at the University of Pennsylvania by Professor Barbara Weber, MD.

REFERENCES

1. Schroeder, HW. Human Heredity. In: Goldman L, Claude J, editors. Cecil textbook of medicine. 21st ed. Philadelphia: W.B. Saunders; 2000. p. 126-159
2. Progeny. [computer program]. Version 2000. South Bend (IN): Progeny Software LLC; 2000. Available from: URL: <http://www.progeny2000.com>
3. Cyrillic. [computer program]. Version 3. Acton(MA): FamilyGenetix Ltd;2000. Available from: URL: <http://www.cyrillicsoftware.com>
4. Brun-Samarq L., Gallina S., Philipppi A., Demenais F., Vaysseix G., Barillot E. CoPE: a collaborative pedigree drawing environment. *Bioinformatics* 1999;15(4):345-6
5. Oracle Clinical. [computer program]. Version 4i.:Oracle Corporation; 2001. Available from: URL: <http://www.oracle.com>
6. ClinTrial. [computer program]. Version 200.:ClinTrials Research Inc.;2001. Available from: URL: <http://www.clintrialsresearch.com>
7. Nadkarni PM, Brandt C, Frawley S, Sayward FG, Einbinder R, Zelterman D, et al. Managing Attribute-Value Clinical Trials Data Using the ACT/DB Client-Server Database System. *J AM Med Inform Assoc.* 1998;5:139-151.
8. Nadkarni PM, Brandt CM, Marengo L. WebEAV: automatic meta-driven generation of web interfaces to entity-attribute-value databases. *J AM Med Inform Assoc.* 2000;7(4):343-56.
9. W3C, Scalable Vector Graphics Specification [HTML document]. Version 1.0.: W3C;2000. Available from: URL: <http://www.w3.org/TR/2000/03/WD-SVG-20000303/index.html>