RECORD LINKING AND GENETIC ANALYSIS OF UTAH DEATH CERTIFICATES

by

Kerry Gene Rowe

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Medical Informatics

The University of Utah

December 1996

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a dissertation submitted by

Kerry G. Rowe

This dissertation has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

[redacted]

Chair: Mark Skolnick

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

FINAL READING APPROVAL

To the Graduate Council of the University of Utah:

I have read the dissertation of [redacted] G. Rowe in its final form and have found that (1) its format, citations, and bibliographic style are consistent and acceptable; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the supervisory committee and is ready for submission to The Graduate School.

[redacted]

[redacted]

[redacted]

Approved for the Major Department

[redacted]

Reed Gardner
Chair/Dean

Approved for the Graduate Council

[redacted]

Ann W. [redacted]
Dean of The Graduate School

ABSTRACT

This research examined the process of probabilistic record linking and the study of familiality for common causes of death. Genealogical records from the Utah Population Database (UPDB) were linked with records from the Utah Cancer Registry and death certificates from the state of Utah. Record linking was done with commercial record linking software from Matchware Technologies called Automatch. The cancer records linked by Automatch were compared to previous linking results to test the effectiveness of the software. The record linking methods established in this linking process were applied to the death certificates.

Death certificate records from 1957 through 1992 were linked to the genealogy records; 126,085 (45%) death certificates were linked. The linked records were grouped by cause of death into the 61 most common causes of death. The familial predisposition of the common causes of death was studied using two different methods.

The first method used the Genealogical Index of Familiality(GIF). The GIF uses the kinship coefficient to measure the degree of relationship between all pairs of individuals in the cause of death groups. The GIF was calculated for each cause of death group along with 100 sets of matching controls. The control mean was compared to the GIF value for the cause of death. Results of the GIF showed substantial familiality for a number of causes of death including kidney cancer, aneurysm, and prostate cancer.

The second method used to examine familiality calculated a first-degree relative risk for the cause of death groups. An expected value for each cause of death was calculated using the resources of the UPDB and compared to the rate of each cause of death among first-degree relatives. A high first-degree relative risk was seen for a number of causes of death including alcohol related deaths, ovarian cancer, and lung cancer.

The record linking software performed well and helped to combine the different data sets successfully. The linked death certificate records provided a large data set to use in the study of familiality for common causes of death.

for Mom and Dad

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

Many individuals assisted me with this endeavor. I would first like to thank Mark Skolnick for serving as my committee chair and making it possible for me complete this research. You were of great assistance in defining my research project and providing timely suggestions to keep it moving towards completion. I truly appreciate your time spent with my research and the efforts you made to insure my financial support.

I would like to acknowledge Lisa Cannon-Allbright for her guidance and encouragement. Your confidence in me and my abilities means a great deal to me. I would also like to especially thank Cathryn Lewis for her help with my statistical problems and the meticulous editing of my written work. Special thanks should also be given to David Goldgar for his help with the relative risk program and methods.

I would like to thank Geri Mineau for all of the time she spent preparing the Utah death certificate records for my use. Your explanations and clarifications of the UPDB and record linking tips were a valuable resource for my research.

I would like to acknowledge the staff of the Department of Genetic Epidemiology for their support with computer problems, programming solutions, and the typing of the numerous graduate forms. I would like to especially thank

Jean Geisler for her friendship and words of encouragement when it seemed that this research would never be completed.

I would like to thank my wife Susie and my children for being patient, understanding, and supportive of me. I know there were many late nights spent at work rather than at home. I am also grateful to my brother Keven for agreeing to edit my dissertation and making helpful suggestions.

This research was supported by NIH grant number CA-64477.

Support for the UPDB is provided by the Pedigree and Population Research Resource, a core institute of the Huntsman Cancer Institute at the University of Utah. Support for computer costs for the UPDB access and record linking work were provided by the Huntsman Cancer Institute at the University of Utah.

CHAPTER 1

INTRODUCTION

The focus of this research is the linking of records from medical and genealogical sources and the use of the linked records to examine familial predisposition for common causes of death. Many medical, genetic, and demographic applications show the value of combining data from multiple sources. Some examples include vital statistics matching, cancer cohort studies, public health and injury surveillance, criminal behavior studies, and highway safety studies. A wealth of genealogical data in the Utah has provided an excellent resource for genetic research, and record linking has played an important part in this research.

The University of Utah maintains a database that includes a genealogical record of the descendants of the Utah pioneers (Skolnick 1979). The pioneers settled in Utah in the middle of the 1800s. The genealogical records database is the foundation of the Utah Population Database (UPDB). The database also contains a cancer registry of all individuals in the state who have had cancer. It was started in 1958 and was statewide by 1966. The State of Utah has also given the university access to the death certificates of individuals who died in Utah between 1957 and 1992. The cancer records have been linked and used in numerous studies that produced significant findings of genetic predispositions to breast, colon, melanoma and prostate cancers (Mikki 1994; Cannon-Albright 1992; Weaver-Feldhaus 1994; McWhorter 1992; Cannon-Albright 1988).

Commercial record linking software, using probabilistic record linking methods, will be used to link records from the genealogical database with the records of the cancer registry and the Utah death certificates. The first objective of this study is to see if the linking software will provide a useful tool for record linkage with the genealogy records. Efforts will be made to obtain the most links possible, while insuring that the links are accurate. The results of the record linking of the cancer registry data will be compared to previous work. This will be done to evaluate the record linking software. The procedures established by linking the cancer registry data will be used to link the death certificates.

Another objective is to demonstrate the usefulness of record linking. This will be done with a comprehensive examination of the familial predispositions for common causes of death. The results of this study will be compared to current research.

1.1 Record Linking

Record linking means bringing together information from two independent sources about the same person. Computer technology has made it possible to combine large numbers of records with relative ease. One of the first demographic and population studies that used record linking with large sets of records was done in a study of the population of Geneva (Henry 1956). A population genetics study of an Italian community was made possible by linking baptismal, marriage, and burial records of Roman Catholic parishes (Skolnick 1971). The linking of these records created a genealogy of the Parma Valley in Italy that dated back to the 1500s. Record linkage has also been used in a number of areas such as medical studies (Cohen 1988), epidemiological studies (Newcombe 1983), database cleansing (Buehler 1989), and data quality (Roos 1989).

The principal steps in record linking are searching for potentially linkable records and comparing the records to determine if they relate to the same person. In the searching step, the aim is to reduce the number of records to be compared and optimize the matching procedure. The matching step takes potentially linkable records and compares them to other records. Linking takes correctly matched records and assembles a composite record for one individual.

To reduce the number of records to be compared, the two files of records can be partitioned into mutually exclusive blocks designed to increase the proportion of matched pairs while decreasing the number of record pairs to compare. The files are sorted by such fields as birth year, last name, or first initial into subsets called blocks. Blocking causes all records having the same value in the blocking fields to be compared. Since the records that are not included in the block are classified as non-matches, an error in a field such as last name or birth year would cause a possible match to be missed. To resolve this problem, multiple matching passes can be run with different blocking parameters.

Although last names are frequently used as blocking variables, they are not particularly efficient, since they are often misspelled or altered. The name's reliability can be improved by setting aside the more unreliable components of the name by using a phonetic coding method. The two most common methods are the Russell Soundex code and the NYSIIS code (Newcombe 1988). The Soundex code is a phonetic coding based on the assignment of code digits which are the same for any of a phonetically similar group of consonants. All of the vowels in a name are discarded. For example Smith and Smythe would both be given the same code of S530. The Soundex system is designed primarily for Anglo-Saxon names but works well with names from a number of different origins. It is not as effective with names of oriental origin, because much of the discriminating power of these names are in the vowels, which are discarded. The NYSIIS codes is similar to the Soundex, except it does retain the position of vowels by changing them all to the letter 'a'. It is more precise than the Soundex method, but both methods work well as blocking parameters. The use of these coding schemes improves the reliability of a name as a blocking variable. Methods used for blocking and searching depend on methods used for matching.

The goal of the matching step is to have the computer apply analytical rules to the records to determine if they do match. The computer attempts to replicate what a human would do (Baldwin 1987). The amount of information available from each record source determines how the records can be linked. In information-rich situations in which large numbers of variables or powerful ones (name, address, social security number) are available, a simple matching or deterministic scheme may be used. Deterministic matching generates links based on the number of agreements among individual identifiers from the two record sources and works best when there are data of high quality (Roos 1991). The main requirement for deterministic matching is that there is an identifier for each individual that is fixed, unique, reliable, and available for every record. There are very few identifiers such as this available in most data registries.

Another approach to deterministic matching is an automated comparison on a character by character basis. This approach usually does not produce satisfactory results. Error rates are high because there are high levels of errors in spelling and recording of names, but errors also occur because the recording of names may vary. A first name for an individual may be recorded as a middle name on another record or as an initial. Nicknames, contractions of names, and the changes of women's names upon marriage produce more errors. Character by character matching is not recommended when accurate matching is required (Gill 1993).

Since there are often unreliable or missing data in record sets, another approach to deciding matches, probabilistic linking, has been developed. Probabilistic linking calculates a weight, based on how closely the fields match, for each field of a record. The weights are added together to produce a score, which will indicate if the records should be matched. A likelihood of correct linkage is calculated. Probabilistic linking methods are useful for data sources where there are missing data or errors. Deterministic linking would not be as accurate in such situations. Probabilistic linking principles are summarized as follows:

> Agreements of various identifying items will generally argue in favor of a linkage, whereas disagreements will argue that the records relate to different people. Numerical weight can be used to quantify the fact that rare names, rare birthplaces, and such, carry more discriminating power when they agree than do their common counterparts. Logarithms are used so the weights can be added and a total weight is achieved at the end indicating the probability the two records do or do not match.
> (Smith 1984, 53)

Weights are used to measure the contribution of each field to the probability of two records matching. An early study discussed the concept of weights based on the probabilities of chance agreement of component value states (Newcombe 1962). A component refers to the fields that are being matched. Another method applied a maximum likelihood method to records that were ambiguous, because of missing information or common names (Skolnick 1973). The distribution of names in the population being linked was used to calculate the probability of compatible matches. The chance of a random match was estimated according to the frequency of a name and the type of record.

The process of record linking was extended with another mathematical application (Fellegi 1969). The definition of weights from this paper takes into account the error probabilities for each field by using a log-likelihood ratio. Each field has two probabilities associated with it called the *m* and *u* probabilities. The *m* probability is the probability that a field agrees given that the record pair being examined is a matched pair. This is effectively one minus the error rate of the field. It is a measurement of the accuracy and reliability of a record field. Deterministic matching can be used to estimate the *m* probability by producing matching records. The fields of the matching pairs can be analyzed to determine the frequency of accurate matching. Fields of matches that were often in error are assigned low *m* probabilities.

The *u* probability is the probability that the record pair being examined is an unmatched pair. Since there are so many more unmatched pairs than matched pairs, this probability is effectively the probability that a field agrees at random. The *u* probability is measurement of the frequency of values for each field.

The *m* and *u* probabilities are used to calculate a weight for each field that is applied if the fields match. If for a given record pair, component *i* matches, the weight for component *i* would be equal to $\log_2(m_i/u_i)$. If component *i* disagrees, then the weight for *i* is equal to $\log_2((1-m_i)/(1-u_i))$. The weights of each field are then added together to compute a composite weight for the two records. On a logarithmic scale, the numerical weight is an estimate of the odds that the two records under consideration do refer to the same individual as opposed to referring to different individuals.

Another concept introduced by Fellegi was an optimal decision procedure for record linkage (Fellegi 1969). For this procedure, three states are defined. A record pair is classified as a match if the composite weight is above a threshold value, a nonmatch if the composite value is below another threshold value, and an undecided situation if the composite weight is between these two thresholds. The setting of these thresholds is a difficult step in probabilistic matching and is accomplished with examination of the record pairs and experimentation.

Further work extended the concepts of record linkage theory by developing a linear sum assignment approach to matching (Jaro 1989). This theory uses an assignment scheme that maximizes the sum of the composite weights of the assigned record pairs and insures that only one record from the first file is matched to only one record from the second file. Also included in this research was the use of the estimation of maximum likelihood model for the *m* probability estimation (Dempster 1977). It uses logarithmic calculations to estimate the error rate in fields where there are incomplete data, which is often the case in record linking. Jaro has compiled his work on record linking into a software product known as Automatch. It is marketed by Matchware Technologies of Burtonsville, Maryland.

Based on a recommendation from a group at Primary Children's Hospital in Salt Lake City, Automatch was purchased to link the genealogy records from the UPDB with the Utah Cancer Registry and Utah death certificates. Although there are a large amount of data in these databases, many of the records are incomplete. Important fields for linking such as birth year and maiden name are often missing. For these reasons, it was determined that a probabilistic linking method should be used with these records. Since Automatch incorporates probabilistic linking methods and can be applied to different types of data, it seemed to be a good choice to use in this record linking project. A successful linking project with Automatch will demonstrate its usefulness in linking the cancer registry records and the death certificates. This will eliminate the need for specialized record linking programs.

1.2 Suitability of Study Population for Genetic Research

The Utah population database was created from the genealogical records of Utah immigrants. The family records that contained at least one individual who was born or died in Utah were collected and used to construct the database. The database consists of over 1,000,000 individuals with 20,000 family founders. Its size insures that common diseases are represented with samples large enough to give significant results. The gene pool is representative of a Northern European population, with low levels of inbreeding and little genetic drift (Skolnick 1987). A major asset of the population for genetic studies is the historical and present desire for large families (Mineau 1979). The large families create the opportunity to study large pedigrees which add significant information to genetic studies. The large pedigrees help to avoid confounding heterogeneity with complex modes of inheritance (Skolnick 1987).

The genealogical records were compiled by members of The Church of Jesus Christ of Latter Day Saints (LDS church). Incidence of some diseases in the Utah population is influenced by the teachings of the LDS church. The members of the church are taught to abstain from alcohol and tobacco use which can be risk factors for developing some diseases. This can be helpful in genetic studies as many of the potential environmental factors for diseases are removed from the analyses.

1.3 Linkage of the Genealogical Records and Medical Databases

The genealogical records are useful in genetic studies of fertility, twinning, and longevity since these phenotypes are part of the genealogical record. The value of the genealogy records in medical genetic studies is greatly increased when they are linked to medical databases. The study of familiality of common diseases is most powerful when ascertainment of all cases in a well-defined population is complete and the genetic relationships between all cases are known (Bishop 1984). This is made possible with the combination of medical and genealogical databases.

Record linking of the genealogy records and the Utah cancer registry has been valuable in cancer genetic studies. Important discoveries in cancer research have been made in cancer research using the linked cancer registry records. Family information found with the linked cancer records have aided in the cloning of the breast cancer gene BRCA1 (Mikki 1994) and the localization of a second breast cancer gene BRCA2 (Wooster 1994). The assignment of a locus of a melanoma susceptibility gene was determined with linked records (Cannon-Albright 1992). The gene was cloned two years later (Weaver-Feldhaus 1994).

Other studies that have shown significant genetic predisposition have been done with prostate cancer (McWhorter 1992) and colorectal cancer (Cannon-Albright 1988) .

Two other studies that examined the familial aggregation of cancers were done using the linked cancer records (Cannon-Albright 1994; Goldgar 1994). One study looked at the genealogical index of familiality, and the other looked at the risk to first degree relatives. Both of these methods involve looking at the frequency of a disease among the relatives of individuals with the disease.

The methods of analysis using the relative risk and genealogical index will be applied to the linked death certificates. A detailed explanation of the methods used for these studies is included with the results of the analysis of the linked death certificates. The findings will be examined to look at familial tendencies for each cause of death.

1.4 Genetic Epidemiological Analysis

A standard method for detecting familiality for a disease is to look for increased rates of the disease among the relatives of affected individuals. The frequency of the disease among relatives is compared to control populations or expected frequencies calculated from population rates. This value is known as a relative risk. The relative risk for cause of death obtained from the death certificates will be calculated for the first-degree relatives of the death cases. A first-degree relative is a parent, sibling, or offspring. The rates of each cause of death for the population of all the people who have died in Utah will be compared with the rates among the first-degree relatives to calculate a relative risk.

The relative risk is a common epidemiological method that is used to determine the risk of disease due to family history, environment or other factors. A number of studies using relative risk have established that there is a significant elevated risk for cancer among individuals with a family history (Cannon-Albright 1991). A comprehensive study was done on breast cancer with the UPDB that showed significant risk for individuals in families with high rates of breast cancer (Slattery 1993). The resources of the UPDB allowed the researchers to determine the risk for distant relatives such as third- to sixth-degree relatives.

The genealogical index of familiality produces an analysis similar to the relative risk, but it goes beyond first-degree relatives and thus is extended to genetic relationships that have a lower probability of sharing nongenetic risk factors such as common environment or diet. It examines the relationship between all possible pairs of individuals in a group and quantifies the relationship by the degree of relatedness for each pair (Hill 1980; Skolnick 1981). The degree of relatedness is measured by the Malécot coefficient of kinship (Malécot 1948). The coefficient of kinship gives a numerical value that describes the degree of relationship for two individuals. The mean of the coefficient of kinship for all the cases with the same cause of death is compared to the mean kinship for a group of age, sex, and birthplace matched controls. The ability to combine the death certificates with the genealogy makes this analysis possible, since the genealogy provides the family relationships between the cases.

The genealogical index of familiality was developed to analyze the data from the UPDB. It requires that all of the cases are ascertained and the genetic relationships between all of the cases be known. These requirements are met by the UPDB with the combination of genealogical records with the statewide cancer registry and the death certificate records.

The genealogical index of familiality was first used to examine the familial predispositions of cardiovascular disease (Williams 1978). Cancer records were studied to examine the familiality of a number of types of cancer (Skolnick 1981). The cancer record study was repeated with updated cancer records and new links to the genealogical records to again examine the familiality of cancer (Cannon-Albright 1994). The analysis of the kinship demonstrated increased familiality for almost all of the cancer sites. The results suggested a genetic susceptibility to most types of cancer at various levels.

The genealogical index produces significant and informative results. However, when there is a small number of cases, the reliability of the results decreases. The problem of insufficient sample size is common in most epidemiological studies. However the size of the UPDB makes it possible to collect large numbers of cases which helps to overcome the limitations of small samples.

1.5 Genetic Epidemiological Analysis of Death Certificates

The relative risk and genealogical index of familiality will provide a comprehensive examination of familial predispositions for common causes of death. With the use of the death certificates, instead of a registry such as the Utah Cancer Registry, a different view of common diseases will be produced. There are many diseases such as cancer in which the incidence of the disease is different from the mortality of the disease. It is expected that the results from the death certificate analysis will differ from previous analyses for that reason. Certain diseases are more severe and occur at a younger age when they are a result of genetic predispositions (Jorde 1995). Additional insight into these diseases will be given by the analysis of the linked death certificates.

CHAPTER 2

RECORD LINKING MATERIALS

Record linking was done with data from the UPDB. The files from the database that were used included a genealogy of Utah pioneers, the Utah Cancer Registry, and Utah death certificates. The record linking software used was a commercial product from Matchware Technologies called Automatch. Automatch uses probabilistic linking techniques to match records and has a number of features that aid a linking project.

2.1 Utah Population Database

The UPDB is a combination of three data sources. The central component of the database is a genealogy of the Utah pioneers and their descendants, which was created in the mid-1970s (Skolnick 1979). There are approximately 1,600,000 individual records and 180,000 family records in the Utah genealogy database. Many of these individuals are the descendants of the original Utah pioneers who were members of the Church of Jesus Christ of Latter Saints, which is commonly known as the LDS church (Skolnick 1980).

The database contains a genealogy that represents the immigrants of Utah and their Utah descendants. The database also contains marriage and birth information for each individual. This information is used to link the founders and their descendants together into family groups. In some instances the records encompass seven generations. The Utah population also has a number of characteristics that make it a valuable resource for genetic studies, such as large family size and a religious emphasis on genealogy. The genealogy records submitted by the descendants of Utah's founders made the creation of the UPDB possible.

The ability to link together multiple generations into large family pedigrees makes the UPDB a valuable tool for genetic research. The number of individuals in the genealogy is large enough to insure that diseases of interest are represented with sample sizes big enough to give significant results in a genetic study. The value of the database in medical genetic research comes from the inclusion of medical record databases that can be linked to the genealogy, such as the Utah Cancer Registry.

The Utah Cancer Registry was started in 1958 and made statewide in 1966. It includes all cancer cases in Utah excluding basal and squamous carcinomas of the skin. The registry maintains abstracts of clinical records and follow-up information on all cases. There are currently 139,475 entries in the registry representing 129,697 individuals.

Another important medical record database in the UPDB is a set of Utah death certificates. There are 292,241 records representing death certificates from

1957 to 1992 in the database. The records contain the first name, last name, and middle initial of each individual. The middle name was added in the early 1970s. Some of the records also have spouse's name and the names of the individual's parents along with birth dates and death dates. The state did not keep the father's name until 1979 and birth date until 1973 in the computer record, so some of this information has been added to the database from microfilm records of the death certificates. The cause of death is coded by the International Classification of Diseases (ICD). The majority of the records use the ninth revision of the ICD code with some of the older records using the seventh or eighth revisions. The early records have only one cause of death listed, whereas some records from the 1980s also have secondary causes of death on the death certificate.

2.2 Automatch Record Linking Software

Automatch is commercial software written for record linking. It was developed by Matthew Jaro of Matchware Technologies. It is based on work he did while working at the US Census Bureau (Jaro 1989). It is available in DOS, IBM MVS, and UNIX versions. The UNIX version, running on a SUN computer, was used in this research. Automatch uses the probabilistic record-linking method. It has features such as multiple-pass matching, unduplication of files, computer-assisted review procedures, and data management of linking records. It also includes specialized functions for file unduplication and geographic coding.

Automatch is currently being used at the Primary Children's Hospital in Salt Lake City, which is academically affiliated with the University of Utah's medical school. Researchers at the hospital have used Automatch in a study that examined the severity of injuries in auto accidents. Hospital admission records were linked with police accident reports to study the severity of injuries suffered in motor vehicle accidents. It is also used in the blood donation departments of hospitals, which are a part of the Intermountain Health Care hospital group. They use the software to keep the records of blood donations free from duplicate records. Based on the recommendations of individuals involved in these applications, it was decided that Automatch would be tried with the UPDB record linkage.

There are a number of steps in record linking with Automatch. The first step is the preparation of the data files and the data dictionaries that describe the files. The data dictionaries contain a name for each field, the location of the field in the file, and the length of the field. An index of each file is constructed based on the matching parameters. The indexing of the data files decreases the amount of storage space needed for the data files and speeds up the matching programs.

The next step is determining the blocking strategies. Blocking is used to reduce the number of pairs that will be compared in each pass, thus making the program more efficient. Blocking does this by grouping the pairs by a common parameter such as birth year or last name. Only those records with the same birth year or similar names are compared.

Automatch recommends the use of a Soundex code when blocking with names. Soundex is a phonetic code, based on the assignment of code digits which are the same for any of a phonetically similar group of consonants. It was designed primarily for Anglo-Saxon names. The Soundex method includes the following steps (Knuth 1973) :

1. Retain the first letter of the name and drop all occurrences of a, e, h, i, o, u, w, y in other positions.

2. Assign the following numbers to the remaining letters after the first:

b, f, p, v → 1

c, g, j, k, q, s, x, z → 2

d, t → 3

l → 4

m, n → 5

r → 6

3. If two or more letters with the same code were adjacent in the original name (before step 1), omit all but the first.

4. Convert to the form "letter, digit, digit, digit" by adding trailing zeros (if there are less than three digits), or by deleting the extra rightmost digits (if there are more than three)

The Soundex code is ideal for use as a blocking variable, since the Soundex code is selective enough to partition names into a fairly large number of blocks, but not so selective that all possible spelling errors are excluded (Jaro 1994). However it is not useful as a matching parameter, since names that are not the same will often be given the same code.

The final step in record linking with Automatch is the preparation of the matching specifications. This involves determining which variables will be matched, the matching parameters, and the match cutoff values. The matching parameters include the *m* and *u* probabilities, which originally are estimated by the user. An estimation of the *m* probability can be calculated by Automatch after a matching pass is run. The *u* probability is calculated by Automatch for each matching pass.

The *m* probability is defined as the probability that a field agrees given that the record pair being examined is a matched pair. This is effectively one minus the error rate of the field. Fields that are critical to the matching process are given high *m* probabilities. Fields that are not as important or that contain missing information or errors in matching records are assigned low *m* probabilities, since the possibility of error is higher. After a match run is completed, there is a program called **mprob** that reviews the results and provides suggested *m* probabilities for each of the fields.

The *u* probability is defined as the probability that a field agrees given that the record pair being examined is an unmatched pair. This is effectively the probability that a field agrees at random. The matching program creates a frequency distribution of the variables in each field, and the *u* probability of each field is calculated from the results. A weight for each field is calculated using the *m* and *u* probabilities. The weight of a field that matches is equal to the $\log_2(m/u)$, whereas the weight for a nonmatching field is equal to $\log_2(1\text{-}m/1\text{-}u)$.

These weight calculations come from the probabilistic linking theory. The weights of all of the fields are summed to get the final score of a possible matching record pair. The 100 most common occurrences of each variable are given a unique weight, and a standard weight is assigned to the rest. This gives the uncommon variables a higher weight than the more common variables.

Other parameters can be defined, such as prorated numeric comparisons for dates and character-uncertainty comparisons for names. The prorated numeric comparisons allow the user to set how much a numeric value can vary and still be considered a match. For example, in order to compensate for data entry errors, the birth year of a record pair could vary by one or more years and still be scored as an exact match. The number of years it can vary is set by the user and is dependent on the quality of the data.

The character uncertainty option employs an information-theoretic string comparison algorithm. It determines the level of similarity between two strings that are similar but not an exact match. The algorithm assigns a value to the comparison of two strings. The user can set a parameter that determines a cutoff level for this score. A score below 700 would mean that the strings are almost certainly different, whereas a score of 900 would mean an exact match. A score between 700 and 900 would let the matching program assign a weight proportional to the score. Automatch recommends using 700 as this parameter, so that the matching program can use the full capabilities of this algorithm.

A typical entry in the match parameter file would have the names of the fields to be matched, the type of variable, the *m* and *u* probabilities and a prorated value for numeric values or character uncertainty level for character strings.

The next step is to run the matching program, which creates files that contain the matched and nonmatched records. The nonmatched records can be used in subsequent match passes. A histogram of match results is made in order to aid in the establishment of cutoff values for the matching scores. Two cutoff values can be set. The highest will signify definite matches, and the other will define marginal matches. The marginal matches can be examined manually with the clerical review program to determine if the records should be classified as matches.

After the first match run is completed, subsequent matching runs can be done on the remaining set of records. Multiple passes compensate for records not matched due to errors in blocking variables. Blocking errors could be a misspelled name or a birth year that is off by a year or a switch of a person's first and middle names. Multiple passes also allow the user to find matches with different sets of matching parameters. There is a limit of eight passes that can be run on a data set. Reports can be created to examine the matching results for each pass.

After completion of the matching runs, the set of matches can then be extracted from the data files created by each pass to create a file of matching records. The contents of these files can include any of the fields from the two original files.

CHAPTER 3

RECORD LINKING RESULTS

The Automatch software was used in two record-linking projects. The first used the UPDB genealogy database and the Utah cancer registry. These two databases had been previously linked with software that was written specifically for the UPDB. The resulting links have been used in cancer genetics research. The record-linking results of Automatch were compared with the previous linking results in order to determine the effectiveness of Automatch. Automatch found more links than the previous work. The tools provided by Automatch were very useful and easy to use. A set of matching techniques and parameters was also established for use in future record linking.

The second linking project used the genealogy database and the Utah death certificates. The linking methods were based on those methods and parameters determined in the cancer registry linking, with some changes due to additional information in the death certificates. This information included the names of the deceased parents and spouse, which were very useful for linking the records. The percentage of death certificates linked was higher than the percentage of cancer records linked. A file containing the linked records and the cause of death for each individual was created.

3.1 Comparison with Existing Cancer Links

A number of projects have linked Utah cancer registry records with genealogy records. The latest effort was completed in 1994. It used a program called LNX, which was written by Dr. Richard Kerber. LNX is based on the probabilistic linking logic presented in the *Handbook of Record Linkage* (Newcombe 1988). The logic looks at the comparison of identifiers such as name and birth date among linked records and unlinked records. It uses this comparison to generate the ratios of the linkable pairs against the unlinkable pairs. This method is similar to the method used by Automatch.

LNX allows the user to select matching and blocking parameters through a menu-based interface. It uses the NYSIIS code for blocking with names. NYSIIS is similar to the Soundex code, except that it retains information on the sequence of vowels by changing them to the letter 'a', instead of discarding them as Soundex does. The NYSIIS code retains more of the discriminating powers of the name but is keeps more of the unreliable components than the Soundex method (Newcombe 1988). This is partly due to Soundex discarding information on the position of vowels in a name. Both methods work well as blocking parameters.

In order to test the effectiveness of Automatch, a set of cancer records was chosen from the cancer registry data. The set of cancer records was linked to the genealogy records, and the set of matches was compared to the results of the

LNX program. The results were validated by comparing them with the previous work. This process of linking and validating helped to develop a procedure for linking records with Automatch that could be used not only with the cancer records but also for other linking projects.

The cancer records chosen included all of the cancer cases in Utah from 1981 through 1990. The records were separated into sets of male and female cases, so sex was a blocking factor in all of the linkage runs. These records were linked to corresponding sets of male and female genealogy records. There were 413,641 female genealogy records and 369,957 male genealogy records. These were linked with 24,869 female cancer records and 24,598 male cancer records. Marriage names of female individuals were added to the female genealogy records when there was a corresponding entry in the UPDB marriage file. This was done, since in the cancer records, the last name of a female is usually a married name. In the genealogy records, the last name of a female record is a maiden name.

A Soundex code was generated for the male last name, the female last name from the cancer registry, and the female married name from the genealogy. The majority of the females had been married, so the married name was the best blocking variable. The Soundex code was used only as a blocking factor but not as a matching variable, since some names that do not match could be assigned the same code.

The fields used for linking the males were last name, first name, first initial, middle initial, middle name, birth state, birth year, birth month, and birthday. The fields for the females included the last name, first name, first initial, middle initial, middle name, birth state, birth year, birth month, and birthday. The third name and maiden name fields from the cancer records along with the married name field from the genealogy were also used for the females. There was a third name field for the males in the cancer records, but since it was usually empty, it was not useful in linking. Some of the fields listed were not used in every linking run, since the parameters used to find matches were different in each run.

The matching parameters for the first match run were chosen by determining the combination of matching parameters and blocking factors that generated the most matches. The choice of parameters to use on the first pass with both the males and the females was straightforward, with the name and birth date fields matched with their corresponding fields in the two data sets. Since the last name in the genealogy data set was the female's maiden name, using the genealogy last name and maiden name was considered for the first run with the female data. However, the number of matches was substantially less than the number found using the married name, so the married name of the female was used first. The maiden name was often missing in the cancer records, causing the reduced number of matches. The maiden name was useful in subsequent runs for matching female records without married names.

An example of using different blocking factors was found with the males.

By using the blocking factor of the first initial instead of birth year, 200 more matches were found. Most of the additional matches were found by allowing the birth year to vary by one year and still be scored as a match. If birth year had been used as a blocking factor, these records would have not been compared. They would have been found in subsequent passes, but finding as many matches as possible in the first run reduced the number of possible matches. This helped to reduce the number of records that needed to be compared in subsequent passes.

The parameters for the subsequent passes were chosen after a review of the record pairs that did not match. One pattern that was easily seen was the switch of the first and middle names. The maiden name for the females was often listed as the middle name or third name, so this pattern was used as a matching parameter.

A test run of the first match pass was made using estimated *m* probabilities. The *m* probability is defined as the probability that a field agrees given the record pair being examined is a match. It is effectively one minus the error rate of the field. The fields that were important for linking, such as first name and last name, were given high *m* probabilities because it was assumed that the error rate of these fields would be small. The last name and first name fields had to be accurate to generate correct matches.

The Automatch **mprob** program was then run to generate its estimation of the *m* probabilities. The *m* probabilities calculated by the Automatch program were used in another test run of the first pass, and the results from the two test runs were compared. The results of the run using the probabilities calculated by **mprob** were better than results with the estimated *m* probabilities, so the **mprob** probabilities were used in all of the passes. The character-uncertainty comparison was used in matching names. The value for this comparison was set at 700, which allowed the matching program to calculate reduced scores for names that were similar. The birth year, birth month, and birthday were allowed to vary by one and still be classified as a match.

The match cutoff scores were determined by viewing the histograms of the score distribution and setting estimated cutoffs based on them. The possible matches around the cutoff were reviewed manually to make sure that there were no false matches above the cutoff line. On some of the passes a clerical review cutoff score was set to manually review records that had not scored above the match cutoff. The record pairs between the match and clerical cutoff were reviewed manually, and those that matched were marked as matches. The pairs that remained were marked as residuals and were used in subsequent passes. Passes that had unusual matching parameters such as first name matching middle name were given a clerical review cutoff. Since the parameters were unusual, the set of possible matches to review was small.

The parameters of each pass are shown in Figures 1 and 2. The blocking variables for each pass are listed first, followed by the matching variables. The cutoff values are listed last with the match cutoff listed first, followed by the

```
BLOCK1 CHAR SDX SDX
BLOCK1 CHAR FINI FINI
MATCH1 UNCERT LNAME LNAME .99 0.01 700
MATCH1 UNCERT FNAME FNAME .98 0.01 700
MATCH1 UNCERT MNAME MNAME .67 0.01 700
MATCH1 CHAR MINI MINI .91 0.01
MATCH1 CHAR BPL BPL .55 0.01
MATCH1 PRORATED BYEAR BYEAR .99 0.01 1
MATCH1 PRORATED BMONTH BMONTH .98 0.01 1
MATCH1 PRORATED BDAY BDAY  .97 0.01 1
CUTOFF1 29.0  29.0

BLOCK2 CHAR SDX SDX
BLOCK2 NUMERIC BYEAR BYEAR
MATCH2 CHAR LNAME LNAME .99 0.01
MATCH2 UNCERT MNAME FNAME .98 0.01 700
MATCH2 CHAR FINI MINI .95 0.01
MATCH2 CHAR BPL BPL .55 0.01
MATCH2 PRORATED BMONTH BMONTH .98 0.01 1
MATCH2 PRORATED BDAY BDAY  .97 0.01 5
CUTOFF2 26.0  26.0

BLOCK3 CHAR SDX SDX
BLOCK3 NUMERIC BYEAR BYEAR
MATCH3 UNCERT LNAME LNAME .99 0.01 700
MATCH3 UNCERT FNAME FNAME .96 0.01 700
MATCH3 UNCERT MNAME MNAME .65 0.01 700
MATCH3 CHAR MINI MINI .91 0.01
MATCH3 CHAR BPL BPL .55 0.01
MATCH3 PRORATED BMONTH BMONTH .98 0.01 1
MATCH3 PRORATED BDAY BDAY  .97 0.01 1
CUTOFF3 26.0  26.0

BLOCK4 CHAR SDX SDX
BLOCK4 CHAR FINI FINI
MATCH4 UNCERT LNAME LNAME .99 0.01 700
MATCH4 CHAR MINI MINI .91 0.01
MATCH4 CHAR BPL BPL .55 0.01
MATCH4 PRORATED BYEAR BYEAR .99 0.01 1
MATCH4 PRORATED BMONTH BMONTH .98 0.01 1
MATCH4 PRORATED BDAY BDAY  .97 0.01 1
CUTOFF4 30.0  20.0
```

Figure 1   Matching Parameters for Male Cancer Records

```
BLOCK5 CHAR FINI FINI
BLOCK5 CHAR LNAME3 LNAME3
MATCH5 UNCERT LNAME LNAME .99 0.01 700
MATCH5 UNCERT FNAME FNAME .96 0.01 700
MATCH5 CHAR MINI MINI .91 0.01
MATCH5 CHAR BPL BPL .52 0.01
MATCH5 PRORATED BMONTH BMONTH .98 0.01 1
MATCH5 PRORATED BDAY BDAY  .97 0.01 5
MATCH5 PRORATED BYEAR BYEAR .99 0.05 1
CUTOFF5 30.0  20.0

BLOCK6 CHAR SDX SDX
BLOCK6 CHAR FINI FINI
MATCH6 UNCERT LNAME LNAME .99 0.01 700
MATCH6 UNCERT FNAME FNAME .98 0.01 700
MATCH6 UNCERT MNAME MNAME .67 0.01 700
MATCH6 CHAR MINI MINI .91 0.01
MATCH6 CHAR BPL BPL .55 0.01
MATCH6 PRORATED BYEAR BYEAR .99 0.01 1
MATCH6 PRORATED BMONTH BMONTH .98 0.01 1
MATCH6 PRORATED BDAY BDAY  .97 0.01 1
CUTOFF6 29.0  19.0
```

Figure 1 continued

```
BLOCK1 CHAR SDX SDX
BLOCK1 CHAR FINI FINI
MATCH1 UNCERT MARR LNAME .99 0.01 700
MATCH1 UNCERT FNAME FNAME .98 0.01 700
MATCH1 UNCERT LNAME MNAME .42 0.01 700
MATCH1 CHAR LINI MINI .64 0.01
MATCH1 CHAR BPL BPL .49 0.01
MATCH1 PRORATED BYEAR BYEAR .98 0.01 1
MATCH1 PRORATED BMONTH BMONTH .99 0.01 1
MATCH1 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF1 29.05  29.05

BLOCK2 CHAR SDX SDX
BLOCK2 CHAR BYEAR BYEAR
MATCH2 UNCERT MARR LNAME .99 0.01 700
MATCH2 UNCERT FNAME FNAME .98 0.01 700
MATCH2 UNCERT MNAME MNAME .42 0.01 700
MATCH2 CHAR MINI MINI .64 0.01
MATCH2 CHAR BPL BPL .55 0.01
MATCH2 PRORATED BMONTH BMONTH .99 0.01 1
MATCH2 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF2 25.0  25.0

BLOCK3 CHAR LNAME3 LNAME3
BLOCK3 CHAR BYEAR BYEAR
MATCH3 UNCERT LNAME LNAME .99 0.01 700
MATCH3 UNCERT FNAME FNAME .98 0.01 700
MATCH3 UNCERT MNAME MNAME .42 0.01 700
MATCH3 CHAR MINI MINI .64 0.01
MATCH3 CHAR BPL BPL .49 0.01
MATCH3 PRORATED BMONTH BMONTH .99 0.01 1
MATCH3 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF3 24.0  20.0

BLOCK4 CHAR SDX SDX
BLOCK4 CHAR BYEAR BYEAR
MATCH4 UNCERT MARR LNAME .99 0.01 700
MATCH4 UNCERT FNAME MNAME .95 0.01 700
MATCH4 UNCERT MNAME FNAME .95 0.01 700
MATCH4 CHAR MINI FINI .64 0.01
MATCH4 CHAR BPL BPL .55 0.01
MATCH4 PRORATED BMONTH BMONTH .99 0.01 1
MATCH4 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF4 30.0  21.0
```

Figure 2   Matching Parameters for Female Cancer Records

```
BLOCK5 CHAR LNAME3 TNAME3
BLOCK5 CHAR FINI FINI
MATCH5 UNCERT LNAME TNAME .99 0.01 700
MATCH5 UNCERT FNAME FNAME .98 0.01 700
MATCH5 UNCERT MNAME MNAME .42 0.01 700
MATCH5 CHAR BPL BPL .49 0.01
MATCH5 PRORATED BMONTH BMONTH .99 0.01 1
MATCH5 PRORATED BDAY BDAY  .98 0.01 1
MATCH5 PRORATED BYEAR BYEAR .98 0.01 1
CUTOFF5 40.0  20.0

BLOCK6 CHAR LNAME3 MAID3
BLOCK6 CHAR FINI FINI
MATCH6 UNCERT LNAME MAID .99 0.01 700
MATCH6 UNCERT FNAME FNAME .98 0.01 700
MATCH6 UNCERT MNAME MNAME .42 0.01 700
MATCH6 CHAR BPL BPL .49 0.01
MATCH6 PRORATED BMONTH BMONTH .99 0.01 1
MATCH6 PRORATED BDAY BDAY  .98 0.01 1
MATCH6 PRORATED BYEAR BYEAR .98 0.01 1
CUTOFF6 35.0  25.0

BLOCK7 NUMERIC BYEAR BYEAR
BLOCK7 CHAR FINI FINI
MATCH7 UNCERT LNAME MNAME .99 0.01 700
MATCH7 UNCERT FNAME FNAME .98 0.01 700
MATCH7 CHAR BPL BPL .49 0.01
MATCH7 PRORATED BMONTH BMONTH .99 0.01 1
MATCH7 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF7 30.0  20.0

BLOCK8 CHAR SDX SDX
BLOCK8 CHAR FINI FINI
MATCH8 UNCERT MARR LNAME .99 0.01 700
MATCH8 UNCERT FNAME FNAME .98 0.01 700
MATCH8 UNCERT MNAME MNAME .42 0.01 700
MATCH8 CHAR MINI MINI .64 0.01
MATCH8 CHAR BPL BPL .55 0.01
MATCH8 PRORATED BMONTH BMONTH .99 0.01 1
MATCH8 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF8 32.0  20.0
```

Figure 2 continued

clerical cutoff. Where both cutoff values are equal, no clerical review was done. A clerical review was not done on the first pass, since the subsequent passes found many of the matches that would have been found with a clerical review. On the last pass of both the male and female sets, a large gap between the match cutoff and clerical cutoff was used. The first pass was then repeated. The clerical review was done to look manually for any matches that may have been missed. The multiple passes reduced the number of possible matches remaining and made a large clerical review on the last pass easier.

A graphical representation of the score distribution for each pass is shown in Figures 3 and 4. These graphs show the frequency of the scores on a log scale. A log scale was used, because the number of nonmatching pairs greatly outnumbered the number of matched pairs. Each cancer record in a block was compared to all of the UPDB records in the block where only one record could match.

The match cutoffs and clerical cutoffs for each pass are shown on each graph. A good indicator of where the cutoff should be set is the point on the graph where the match and nonmatch line dips to zero. This was usually the place where the number of nonmatches decreased. The cutoff values were different for each pass, since different blocking factors and matching parameters were used. The record pairs to the left of the match line or clerical line are non-matches, whereas those to the right of the match line were classified as matches. Where a clerical review was done, the clerical line was drawn. The record pairs

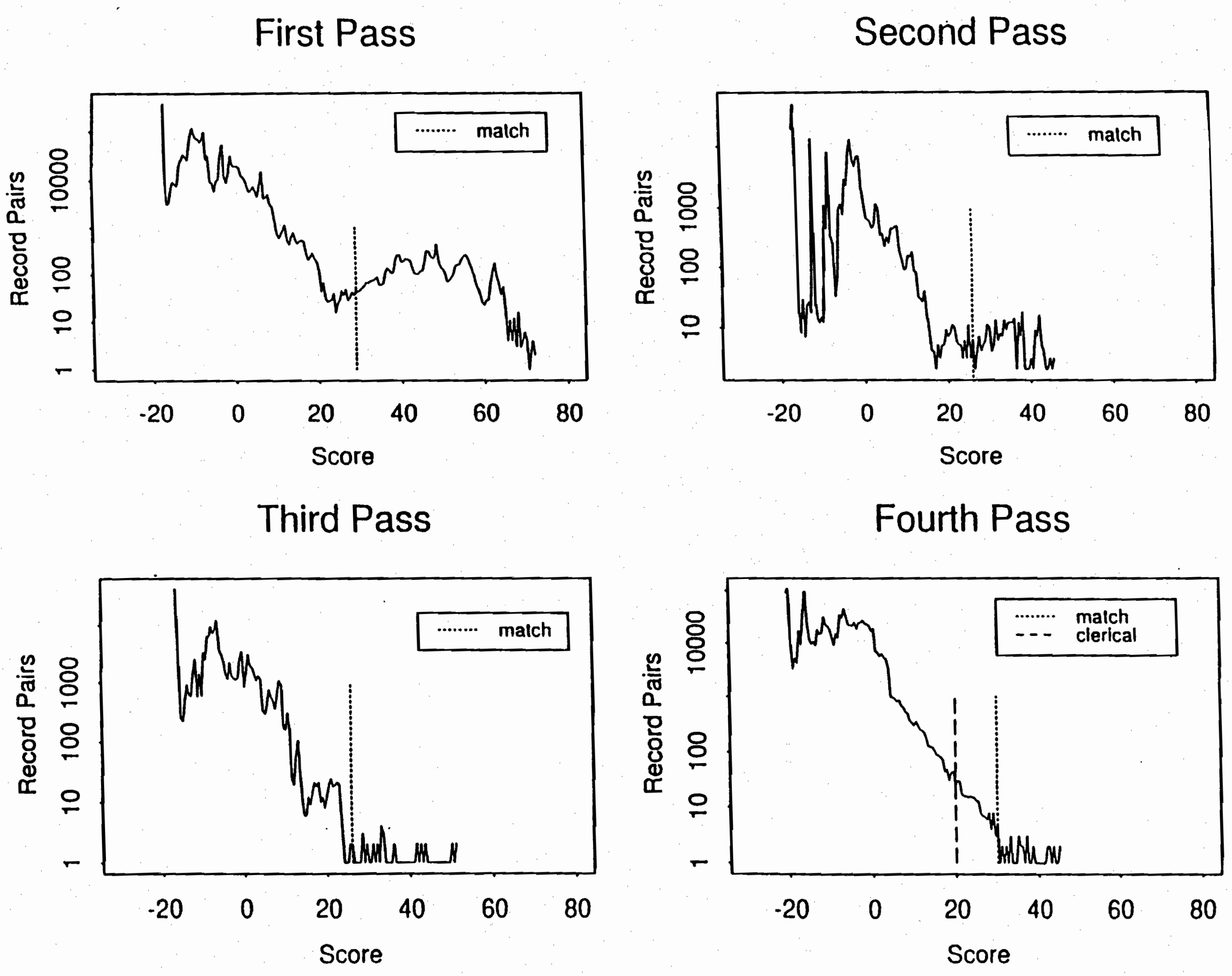


Figure 3 Male Cancer Matching Histograms

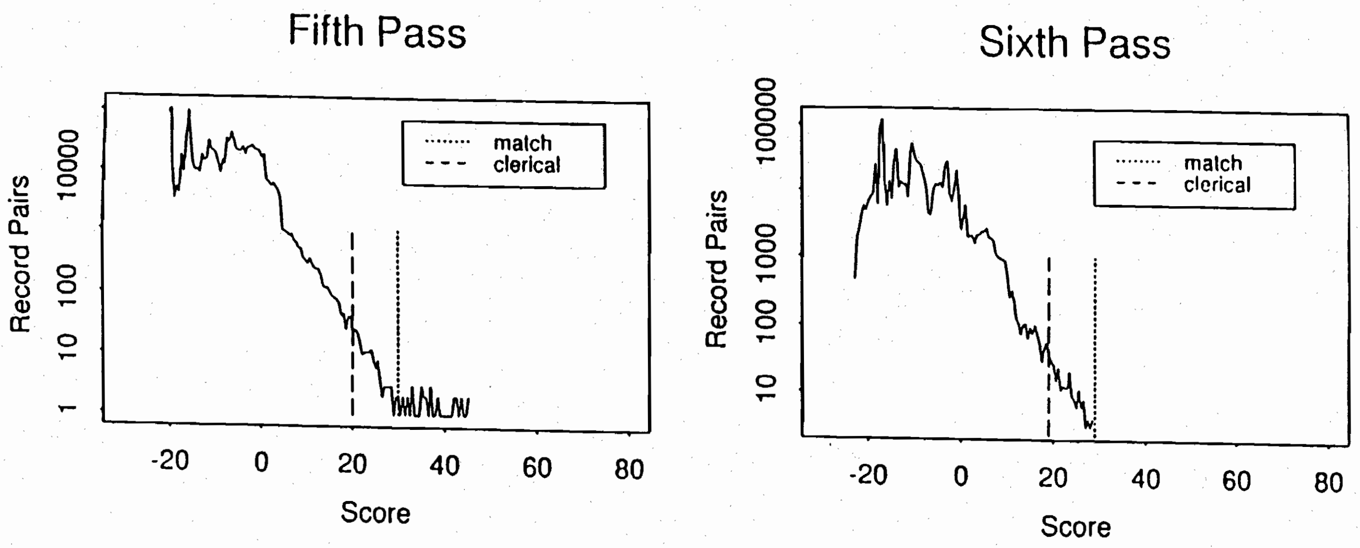


Figure 3 continued

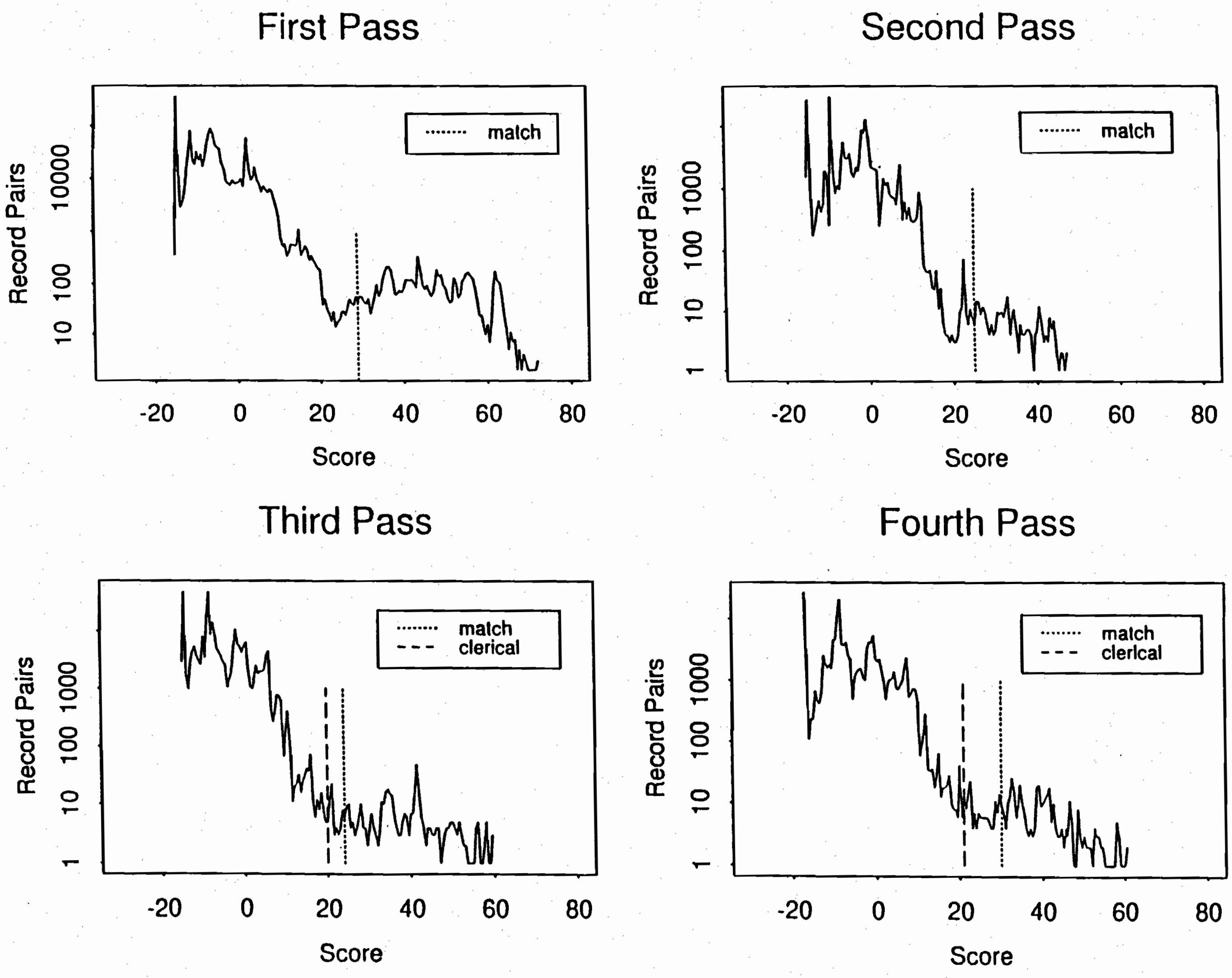
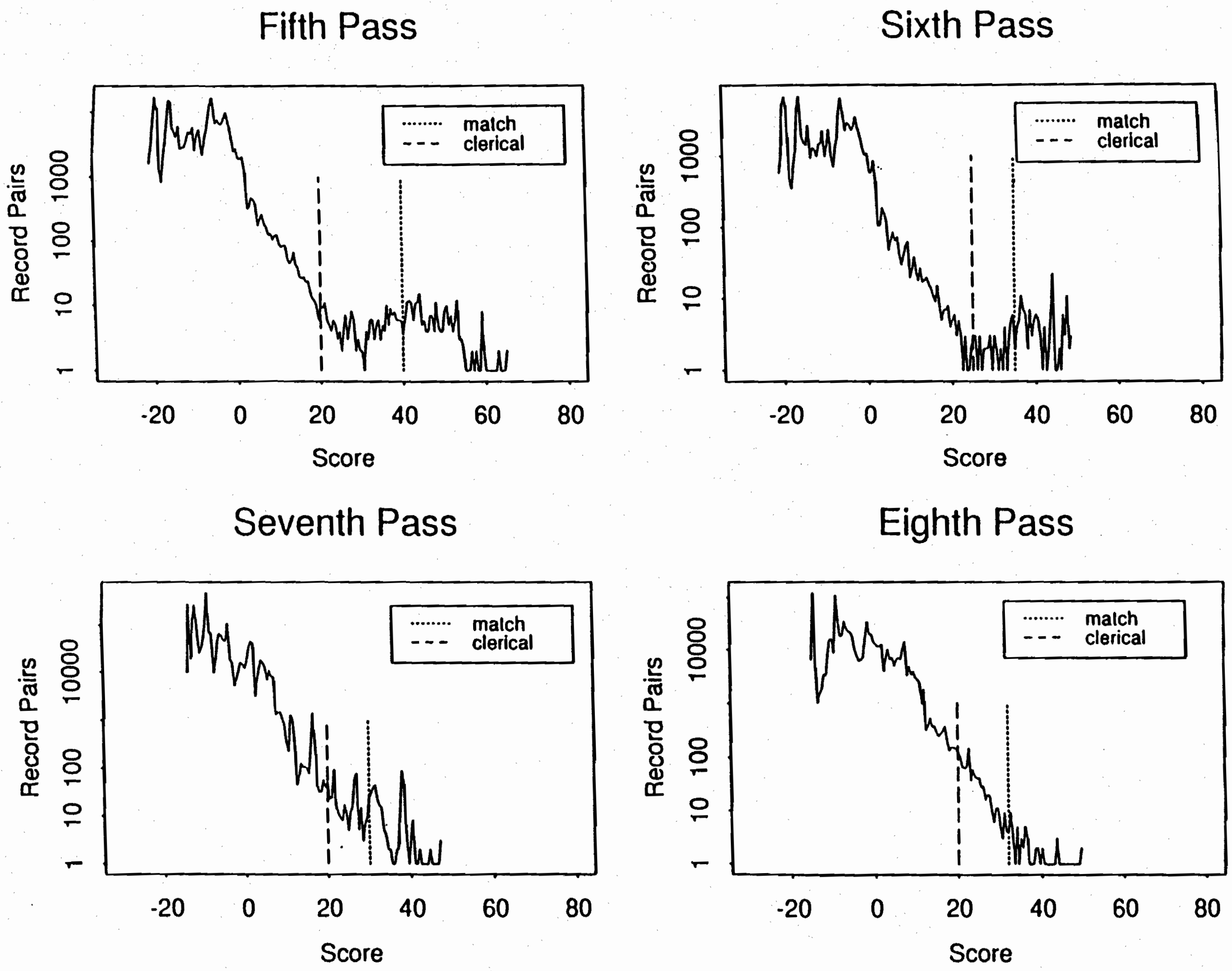


Figure 4 Female Cancer Matching Histograms



Fifth Pass

Sixth Pass

Seventh Pass

Eighth Pass

Figure 4 continued

between the two lines were reviewed manually. The graphs show how the number of possible matches decreased with each pass.

The number of matches found decreased greatly after the first pass, which is expected from the matching design. The number of male records matched decreases sequentially with each pass, whereas the females did not. This is due to the many different combinations of female name matching parameters used. Eight passes were done on the female set and only six passes done on the male set. There were not as many combinations of names to try with the male set when compared with the female set. Other combinations of names that could have been used with the females, but Automatch limits the number of passes that can be run to eight. Because of this limit, the passes that generated the most matches were run. Table 1 shows the number of matches that were found in each pass.

Table 1 Matches Found by Pass Number

Males		Females	
Pass	Matches found	Pass	Matches found
1	9,251	1	6,216
2	243	2	230
3	110	3	354
4	51	4	383
5	40	5	240
6	18	6	121
		7	128
		8	74

The table shows that with the male data set, 96% of the matches were found with the first pass. In the female set only 76% of the matches were found with the first pass. Some of the females were either not married or were married multiple times and had two different married names. Since the first pass on the female set used married name as a blocking factor, a mismatch in the married name field would not let the records be compared, thus producing the reduced number of matches. These matches were found in subsequent passes using different blocking factors.

Special consideration was given to individuals who were identified as twins in the genealogy records, since twins have the same birth date and often have similar first names. All of the information for any twin that was linked was retrieved and reviewed manually with no errors found. The lack of errors is a result of the identification of duplicate records from the data files. If Automatch found a duplicate pair in the genealogy that linked to a cancer record, it would assign the match to the genealogy record that had the highest score with the cancer record. Thus if a twin was linked, the first name that matched the best would receive the highest score and the correct twin was matched with the cancer record.

The results of the Automatch record linking were compared to the existing links found by the LNX program. The comparison is shown in Table 2. Automatch linked 9,713 (39.5%) of the male records, compared to 9,250 (37.6%) linked by LNX. Automatch linked 8,189 (32.9%) of the female records, compared

Table 2 Automatch Record Linking vs LNX Record Linking

	Males	Females
Total cancer records	24,598	24,869
Automatch links	9,713 (39.5%)	8,189 (32.9%)
Lnx links	9,250 (37.6%)	7,663 (30.8%
Linked by both	9,214	7,569
Only LNX	36	93
Only Automatch	499	620

to 7,663 (30.8%) linked by LNX. There were 499 male matches and 620 female matches found only by Automatch, and there were 36 male matches and 93 female matches found only by LNX.

The records that were linked only by Automatch or LNX were examined manually. There were a small number of erroneous matches in both sets. The majority of the errors were in the female data sets. The female data sets were more difficult to link due to the number of changes made with names during the lifetime of the females. The different uses of a maiden name as a middle name or third name also caused difficulty. Also some females used previous married names as middle and third names.

Twenty-three false matches in the Automatch set were deleted from the totals listed in Table 2. There were six false matches in the male set and 17 false matches in the female set. In the male records, five of the records matched exactly on names, birthday, and birth month, but the birth year differed by more than 20 years. These were most likely fathers who had named their son after themselves. One record pair had the same last name, middle name, first initial, birthday, and birth month, but the first name and birth year were different. Eleven of the female records had matching names, but the birth year was not close. There were six female matches where the first and middle names matched, along with the birth date, but the last name or married names did not match.

The records that were linked only by LNX were examined to determine why Automatch did not link them. There were 17 male matches and 10 female matches that were questionable matches. Some of these matches had the birth date matching exactly, but the first and middle names did not match. With other pairs, the names matched, but the birth dates did not match.

One common reason for Automatch missing a valid match was a misspelling of a name. Twelve of the male records and 31 of the female records were missed for this reason. Examples of misspelled names are last names of Toble and Tarvis, instead of Noble and Jarvis. Last name was an important field because it was used for blocking in many of the match passes. It was also weighted heavily so that a match on a last name would receive a high score. If a record pair did not match on a last name, it would not be classified as a match in most cases.

Some of the links were missed by Automatch due to the combination of common names such as Brown or Joseph and the records missing a middle name or birthday. The score for matching a common name was not high enough for the pair to be classified as a match if there was missing data in other fields. Four of the male records were missed for this reason. Another problem was the use of nicknames such as Lori instead of Lorraine. Six of the female matches when nicknames were not matched. Other matches were missed because birthday and birth month were switched or missing.

Most of the matches that were missed with the female data set were due to not matching on the right combination of names. There were nine missed matches where the third name from the cancer data matched the maiden name from the genealogy data, but there was a mismatch in another field. The myriad combinations of names for the female data set made it difficult to get all of the possible matches. There were other missed matches for which important data such as maiden name were missing or there were two different married names.

Overall, the Automatch program performed well and found substantially more matches than the previous work. It worked well with the cancer and genealogy data sets. Setting the cutoff values took the most time and effort. It was usually best to do a small clerical review, since there were often a number of matches mixed with nonmatches just below the match cutoff value. Errors usually occurred when birth year was not used as a blocking factor, so extra attention is needed in this case.

3.2 Death Certificate Linking

Utah joined the Death Registration Act of the federal government in 1912. Microfilm records of death certificates were begun in 1904. The state has computerized records that begin in 1956, although much of the 1956 data is not a part of the computerized record. There are three sets of death certificate data which were given to the University of Utah and made a part of the UPDB. These include death certificates for the years 1957-1979, 1980-1988, and 1989-1992. The state changed the format of the death certificate coding a number of times between 1957 and 1992. Some of the differences in format are the addition of more family information and secondary causes of death in the more recent sets.

The first set of death certificates covering the years from 1956 to 1981 was given to the university in the early 1980s. The amount of information that was in the computerized record was limited. A group led by Dr. Roger Williams was able to add supplementary information to the computerized records from the microfilm records. This information included birth date, middle name, parent's name and spouse's name. Several years were not completed, which made linking records from those years difficult. The years that were not updated are reflected in the linking results.

The death certificates for 1982 to 1992 were given to the university in 1994. The state added birth date in 1973 and father's last name to the computerized record in 1979. In 1989, mother's name, father's first name, and spouse's name were added. The information that was useful in linking for each data set is summarized in Table 3. Other information from the death certificate that was useful in the records analysis included place of birth and the county where the individual died. The state limited the availability of death certificates to those

Table 3 Fields Used in Death Certificate Linking

Death year	1957-79	1980-88	1989-92
	last name	last name	last name
	first name	first name	first name
	middle name	middle name	middle name
	birth date	birth date	birth date
	father's last name	father's last	father's last name
	father's first name		father's first name
	mother's first name		mother's first name
	spouse's first name		spouse's first name

individuals who had died in Utah, which excludes those residents of Utah who died outside the state.

Because the female last name in the genealogy database was her maiden name, the father's last name or maiden name was valuable in the record linking. It was included in the death certificates more than in the cancer records, where it was often missing. The father's, mother's, and spouse's first name were useful as an additional attribute for linking records in which there was a questionable match for the individual's name or where there was an error in the birth date.

The approach to linking the death certificate records was the same that was used in the cancer record linking. The first passes used the name fields from the genealogy matching with the corresponding name fields from the death certificates. The additional parameters such as spouse's name and mother's name were also used. These parameters were the most useful in the 1989-92 set of death certificates, because they were not included in the 1980-88 set and were often missing in the 1957-79 set.

Careful attention was paid to birth year by setting high match cutoff values when birth year was used as a blocking variable. This was done to prevent the type of error that occurred in the cancer linking when the names matched exactly and birth month or birthday matched but birth year was not close. This error produced a score that was classified as a match, when it should not have been. When birth year was not used as a blocking factor in the later passes, a high match cutoff was set, along with a low cutoff for the manual review. The manual review allowed a close check of the birth year to insure a correct match.

The matching histograms were used to set the match and clerical cutoff values. The matching parameters and the corresponding histogram graphs for each set are shown in Figures 5 - 16. The graphs are similar to the cancer linking graphs. They show that most of the male matches were found in the first passes, as was the case in the cancer record linking. The match values are shown on each graph, along with the clerical cutoff if one was set for the matching pass.

The cutoff values were determined by reviewing the histograms. Some records around the match cutoff were examined to make sure that no false matches were generated. The methods of defining the *m* probabilities were the same as those used in the cancer linking. The birth date fields were allowed to

```
BLOCK1 CHAR SDX SDX
BLOCK1 CHAR BYEAR BYEAR
MATCH1 UNCERT SLAST LNAME .99 0.01 700
MATCH1 UNCERT FNAME FNAME .98 0.01 700
MATCH1 CHAR MINI MINI .25 0.01
MATCH1 PRORATED BMONTH BMONTH .99 0.01 1
MATCH1 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF1 25.0  25.0

BLOCK2 CHAR SDX SDX
BLOCK2 CHAR BYEAR BYEAR
MATCH2 UNCERT FLAST FLAST .98 0.01 700
MATCH2 UNCERT FNAME FNAME .98 0.01 700
MATCH2 UNCERT MINI MINI .25 0.01 700
MATCH2 PRORATED BMONTH BMONTH .99 0.01 1
MATCH2 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF2 25.0  25.0

BLOCK3 CHAR SLAST3 LNAME3
BLOCK3 CHAR BYEAR BYEAR
MATCH3 UNCERT SLAST LNAME .99 0.01 700
MATCH3 UNCERT FNAME FNAME .98 0.01 700
MATCH3 UNCERT LINI MINI .25 0.01 700
MATCH3 PRORATED BMONTH BMONTH .99 0.01 1
MATCH3 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF3 26.0  24.0

BLOCK4 CHAR LNAME3 LNAME3
BLOCK4 CHAR BYEAR BYEAR
MATCH4 UNCERT LNAME LNAME .95 0.01 700
MATCH4 UNCERT FNAME FNAME .98 0.01 700
MATCH4 UNCERT MINI MINI .25 0.01 700
MATCH4 PRORATED BMONTH BMONTH .99 0.01 1
MATCH4 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF4 25.0  22.0
```

Figure 5 Matching Parameters for Female Records 1957-1979

```
BLOCK5 CHAR SLAST3 LNAME3
BLOCK5 CHAR BYEAR BYEAR
MATCH5 UNCERT SLAST LNAME .99 0.01 700
MATCH5 UNCERT FINI MINI .95 0.01 700
MATCH5 UNCERT MINI FINI .95 0.01 700
MATCH5 PRORATED BMONTH BMONTH .99 0.01 1
MATCH5 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF5 32.0  20.0

BLOCK6 CHAR LNAME3 FLAST3
BLOCK6 CHAR FINI FINI
MATCH6 UNCERT LNAME FLAST .99 0.01 700
MATCH6 UNCERT FNAME FNAME .98 0.01 700
MATCH6 PRORATED BMONTH BMONTH .99 0.01 1
MATCH6 PRORATED BDAY BDAY  .98 0.01 1
MATCH6 PRORATED BYEAR BYEAR .98 0.01 1
CUTOFF6 33.0  26.0

BLOCK7 CHAR SLAST3 LNAME3
BLOCK7 CHAR FINI FINI
MATCH7 UNCERT SLAST LNAME .99 0.01 700
MATCH7 UNCERT FNAME FNAME .98 0.01 700
MATCH7 CHAR MINI MINI .33 0.01
MATCH7 PRORATED BYEAR BYEAR .98 0.01 1
MATCH7 PRORATED BMONTH BMONTH .99 0.01 1
MATCH7 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF7 29.0  24.0

BLOCK8 CHAR SLAST3 LNAME3
BLOCK8 CHAR BYEAR BYEAR
MATCH8 UNCERT SLAST LNAME .99 0.01 700
MATCH8 UNCERT FNAME FNAME .98 0.01 700
MATCH8 CHAR MINI MINI .33 0.01
MATCH8 PRORATED BMONTH BMONTH .99 0.01 1
MATCH8 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF8 27.0  20.0
```

Figure 5 continued

```
BLOCK1 CHAR SDX SDX
BLOCK1 CHAR BYEAR BYEAR
MATCH1 UNCERT LNAME LNAME .99 0.01 700
MATCH1 UNCERT FNAME FNAME .99 0.01 700
MATCH1 CHAR MINI MINI .96 0.01
MATCH1 PRORATED BMONTH BMONTH .99 0.01 1
MATCH1 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF1 24.0  24.0

BLOCK2 CHAR SDX SDX
BLOCK2 NUMERIC FINI FINI
MATCH2 CHAR LNAME LNAME .99 0.01
MATCH2 UNCERT FNAME FNAME .99 0.01 700
MATCH2 CHAR MINI MINI .96 0.01
MATCH2 PRORATED BYEAR BYEAR .98 0.01 1
MATCH2 PRORATED BMONTH BMONTH .99 0.01 1
MATCH2 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF2 30.0  30.0

BLOCK3 CHAR SDX SDX
BLOCK3 NUMERIC BYEAR BYEAR
MATCH3 UNCERT LNAME LNAME .99 0.01 700
MATCH3 CHAR MNAME FNAME .96 0.01
MATCH3 CHAR SFIRST SFIRST .75 0.01
MATCH3 PRORATED BMONTH BMONTH .99 0.01 1
MATCH3 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF2 30.0  22.0

BLOCK4 CHAR SDX SDX
BLOCK4 CHAR FINI FINI
MATCH4 UNCERT LNAME LNAME .99 0.01 700
MATCH4 UNCERT FNAME FNAME .99 0.01 700
MATCH4 CHAR MINI MINI .96 0.01
MATCH4 CHAR SFIRST SFIRST .77 0.01
MATCH4 PRORATED BYEAR BYEAR .96 0.01 1
MATCH4 PRORATED BMONTH BMONTH .98 0.01 1
MATCH4 PRORATED BDAY BDAY  .97 0.01 1
CUTOFF4 35.5  30.5
```

Figure 6   Matching Parameters for Male Records 1957-1979

```
BLOCK5 CHAR FINI FINI
BLOCK5 CHAR BYEAR BYEAR
MATCH5 UNCERT LNAME LNAME .99 0.01 700
MATCH5 UNCERT FNAME FNAME .99 0.01 700
MATCH5 CHAR MINI MINI .96 0.01
MATCH5 PRORATED BMONTH BMONTH .99 0.01 1
MATCH5 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF5 28.0  22.0

BLOCK6 CHAR FINI FINI
BLOCK6 CHAR LNAME3 LNAME3
MATCH6 PREFIX LNAME LNAME .99 0.01
MATCH6 UNCERT FNAME FNAME .99 0.01 700
MATCH6 CHAR MINI MINI .96 0.01
MATCH6 CHAR SFIRST SFIRST .77 0.01
MATCH6 CHAR MFIRST MFIRST .77 0.01
MATCH6 CHAR FFIRST FFIRST .77 0.01
MATCH6 PRORATED BMONTH BMONTH .99 0.01 1
MATCH6 PRORATED BDAY BDAY  .98 0.01 5
MATCH6 PRORATED BYEAR BYEAR .99 0.05 1
CUTOFF6 33.5  31.0

BLOCK7 CHAR FINI FINI
BLOCK7 CHAR LNAME3 LNAME3
BLOCK7 NUMERIC BMONTH BMONTH
MATCH7 PREFIX LNAME LNAME .99 0.01
MATCH7 UNCERT FNAME FNAME .99 0.01 700
MATCH7 CHAR MINI MINI .96 0.01
MATCH7 CHAR SFIRST SFIRST .77 0.01
MATCH7 PRORATED BMONTH BMONTH .99 0.01 1
MATCH7 PRORATED BDAY BDAY  .98 0.01 5
MATCH7 PRORATED BYEAR BYEAR .99 0.05 1
CUTOFF7 33.5  28.0
```

Figure 6 continued

```
BLOCK1 CHAR SDX SDX
BLOCK1 CHAR FINI FINI
MATCH1 UNCERT MARR LNAME .98 0.01 700
MATCH1 UNCERT FNAME FNAME .98 0.01 700
MATCH1 UNCERT LNAME MNAME .35 0.01 700
MATCH1 CHAR LINI MINI .51 0.01
MATCH1 CHAR BPL BPL .95 0.01
MATCH1 PRORATED BYEAR BYEAR .98 0.01 1
MATCH1 PRORATED BMONTH BMONTH .99 0.01 1
MATCH1 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF1 32.0  32.0

BLOCK2 CHAR SDX SDX
BLOCK2 CHAR BYEAR BYEAR
MATCH2 UNCERT LNAME MAID .99 0.01 700
MATCH2 UNCERT FNAME FNAME .98 0.01 700
MATCH2 UNCERT MNAME MNAME .42 0.01 700
MATCH2 CHAR MINI MINI .64 0.01
MATCH2 CHAR BPL BPL .95 0.01
MATCH2 PRORATED BMONTH BMONTH .99 0.01 1
MATCH2 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF2 25.0  25.0

BLOCK3 CHAR LNAME3 LNAME3
BLOCK3 CHAR BYEAR BYEAR
MATCH3 UNCERT LNAME LNAME .99 0.01 700
MATCH3 UNCERT FNAME FNAME .98 0.01 700
MATCH3 UNCERT MNAME MNAME .42 0.01 700
MATCH3 CHAR MINI MINI .64 0.01
MATCH3 CHAR BPL BPL .95 0.01
MATCH3 PRORATED BMONTH BMONTH .99 0.01 1
MATCH3 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF3 25.0  22.0

BLOCK4 CHAR MARR3 LNAME3
BLOCK4 CHAR BYEAR BYEAR
MATCH4 UNCERT MARR LNAME .99 0.01 700
MATCH4 UNCERT FNAME MNAME .95 0.01 700
MATCH4 UNCERT MNAME FNAME .95 0.01 700
MATCH4 CHAR MINI FINI .64 0.01
MATCH4 CHAR BPL BPL .95 0.01
MATCH4 PRORATED BMONTH BMONTH .99 0.01 1
MATCH4 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF4 30.0  19.0
```

Figure 7   Matching Parameters for Female Records 1980-1988

```
BLOCK5 CHAR LNAME3 MAID3
BLOCK5 CHAR FINI FINI
MATCH5 UNCERT LNAME MAID .99 0.01 700
MATCH5 UNCERT FNAME FNAME .98 0.01 700
MATCH5 UNCERT MNAME MNAME .42 0.01 700
MATCH5 CHAR BPL BPL .95 0.01
MATCH5 PRORATED BMONTH BMONTH .99 0.01 1
MATCH5 PRORATED BDAY BDAY  .98 0.01 1
MATCH5 PRORATED BYEAR BYEAR .98 0.01 1
CUTOFF5 26.0  18.0

BLOCK6 NUMERIC BYEAR BYEAR
BLOCK6 CHAR FINI FINI
MATCH6 UNCERT LNAME MNAME .99 0.01 700
MATCH6 UNCERT FNAME FNAME .98 0.01 700
MATCH6 CHAR BPL BPL .95 0.01
MATCH6 PRORATED BMONTH BMONTH .99 0.01 1
MATCH6 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF6 30.0  17.0

BLOCK7 CHAR SDX SDX
BLOCK7 CHAR FINI FINI
MATCH7 UNCERT MARR LNAME .99 0.01 700
MATCH7 UNCERT FNAME FNAME .98 0.01 700
MATCH7 UNCERT MNAME MNAME .42 0.01 700
MATCH7 CHAR MINI MINI .64 0.01
MATCH7 CHAR BPL BPL .95 0.01
MATCH7 PRORATED BMONTH BMONTH .99 0.01 1
MATCH7 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF7 32.0  20.0
```

Figure 7 continued

```
BLOCK1 CHAR SDX SDX
BLOCK1 CHAR FINI FINI
MATCH1 UNCERT LNAME LNAME .99 0.01 700
MATCH1 UNCERT FNAME FNAME .99 0.01 700
MATCH1 UNCERT MNAME MNAME .71 0.01 700
MATCH1 CHAR MINI MINI .96 0.01
MATCH1 CHAR BPL BPL .96 0.01
MATCH1 PRORATED BYEAR BYEAR .98 0.01 1
MATCH1 PRORATED BMONTH BMONTH .99 0.01 1
MATCH1 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF1 33.8  33.8

BLOCK2 CHAR SDX SDX
BLOCK2 NUMERIC BYEAR BYEAR
MATCH2 CHAR LNAME LNAME .99 0.01
MATCH2 UNCERT MNAME FNAME .99 0.01 700
MATCH2 CHAR FINI MINI .96 0.01
MATCH2 CHAR BPL BPL .96 0.01
MATCH2 PRORATED BMONTH BMONTH .99 0.01 1
MATCH2 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF2 26.0  20.0

BLOCK3 CHAR SDX SDX
BLOCK3 NUMERIC BYEAR BYEAR
MATCH3 UNCERT LNAME LNAME .99 0.01 700
MATCH3 UNCERT FNAME FNAME .99 0.01 700
MATCH3 UNCERT MNAME MNAME .71 0.01 700
MATCH3 CHAR MINI MINI .96 0.01
MATCH3 CHAR BPL BPL .96 0.01
MATCH3 PRORATED BMONTH BMONTH .99 0.01 1
MATCH3 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF3 24.0  24.0

BLOCK4 CHAR SDX SDX
BLOCK4 CHAR FINI FINI
MATCH4 UNCERT LNAME LNAME .99 0.01 700
MATCH4 UNCERT FINI FINI .99 0.01 700
MATCH4 CHAR MINI MINI .96 0.01
MATCH4 CHAR BPL BPL .96 0.01
MATCH4 PRORATED BYEAR BYEAR .99 0.01 1
MATCH4 PRORATED BMONTH BMONTH .99 0.01 1
MATCH4 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF4 40.0  20.0
```

Figure 8   Matching Parameters for Male Records 1980-1988

```
BLOCK5 CHAR FINI FINI
BLOCK5 CHAR LNAME3 LNAME3
MATCH5 UNCERT LNAME LNAME .99 0.01 700
MATCH5 UNCERT FNAME FNAME .99 0.01 700
MATCH5 CHAR MINI MINI .96 0.01
MATCH5 CHAR BPL BPL .96 0.01
MATCH5 PRORATED BMONTH BMONTH .99 0.01 1
MATCH5 PRORATED BDAY BDAY  .98 0.01 5
MATCH5 PRORATED BYEAR BYEAR .99 0.05 1
CUTOFF5 40.0  20.0

BLOCK6 CHAR SDX SDX
BLOCK6 CHAR FINI FINI
MATCH6 UNCERT LNAME LNAME .99 0.01 700
MATCH6 UNCERT FNAME FNAME .98 0.01 700
MATCH6 UNCERT MNAME MNAME .98 0.01 700
MATCH6 CHAR MINI MINI .96 0.01
MATCH6 CHAR BPL BPL .96 0.01
MATCH6 PRORATED BYEAR BYEAR .99 0.01 1
MATCH6 PRORATED BMONTH BMONTH .99 0.01 1
MATCH6 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF6 29.0  17.0
```

Figure 8 continued

```
BLOCK1 CHAR SLAST3 LNAME3
BLOCK1 CHAR FINI FINI
MATCH1 UNCERT SLAST LNAME .98 0.01 700
MATCH1 UNCERT FNAME FNAME .98 0.01 700
MATCH1 UNCERT LNAME MNAME .35 0.01 700
MATCH1 CHAR LINI MINI .51 0.01
MATCH1 PRORATED BYEAR BYEAR .98 0.01 1
MATCH1 PRORATED BMONTH BMONTH .99 0.01 1
MATCH1 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF1 30.15  30.15

BLOCK2 CHAR SDX SDX
BLOCK2 CHAR BYEAR BYEAR
MATCH2 UNCERT LNAME FLAST .99 0.01 700
MATCH2 UNCERT FNAME FNAME .98 0.01 700
MATCH2 UNCERT MNAME MNAME .42 0.01 700
MATCH2 CHAR MINI MINI .64 0.01
MATCH2 PRORATED BMONTH BMONTH .99 0.01 1
MATCH2 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF2 24.0  24.0

BLOCK3 CHAR LNAME3 LNAME3
BLOCK3 CHAR BYEAR BYEAR
MATCH3 UNCERT LNAME LNAME .51 0.01 700
MATCH3 UNCERT FNAME FNAME .98 0.01 700
MATCH3 UNCERT MFIRST MFIRST .90 0.01 700
MATCH3 UNCERT FFIRST FFIRST .90 0.01 700
MATCH3 PRORATED BMONTH BMONTH .99 0.01 1
MATCH3 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF3 25.0  22.0

BLOCK4 CHAR SLAST3 LNAME3
BLOCK4 CHAR BYEAR BYEAR
MATCH4 UNCERT SLAST LNAME .99 0.01 700
MATCH4 UNCERT FNAME MNAME .95 0.01 700
MATCH4 UNCERT MNAME FNAME .95 0.01 700
MATCH4 PRORATED BMONTH BMONTH .99 0.01 1
MATCH4 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF4 30.0  19.0
```

Figure 9 Matching Parameters for Female Records 1989-1992

```
BLOCK5 CHAR LNAME3 FLAST3
BLOCK5 CHAR FINI FINI
MATCH5 UNCERT LNAME FLAST .99 0.01 700
MATCH5 UNCERT MFIRST MFIRST .90 0.01 700
MATCH5 UNCERT FFIRST FFIRST .90 0.01 700
MATCH5 UNCERT FNAME FNAME .98 0.01 700
MATCH5 PRORATED BMONTH BMONTH .99 0.01 1
MATCH5 PRORATED BDAY BDAY  .98 0.01 1
MATCH5 PRORATED BYEAR BYEAR .98 0.01 1
CUTOFF5 32.0  20.0

BLOCK6 CHAR SDX SDX
BLOCK6 CHAR FINI FINI
MATCH6 UNCERT SLAST LNAME .99 0.01 700
MATCH6 UNCERT FNAME FNAME .98 0.01 700
MATCH6 UNCERT MNAME MNAME .42 0.01 700
MATCH6 CHAR MINI MINI .64 0.01
MATCH6 PRORATED BMONTH BMONTH .99 0.01 1
MATCH6 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF6 32.0  20.0
```

Figure 9 continued

```
BLOCK1 CHAR SDX SDX
BLOCK1 CHAR FINI FINI
MATCH1 UNCERT LNAME LNAME .99 0.01 700
MATCH1 UNCERT FNAME FNAME .99 0.01 700
MATCH1 UNCERT MNAME MNAME .71 0.01 700
MATCH1 CHAR MINI MINI .96 0.01
MATCH1 PRORATED BYEAR BYEAR .98 0.01 1
MATCH1 PRORATED BMONTH BMONTH .99 0.01 1
MATCH1 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF1 33.0  33.0

BLOCK2 CHAR SDX SDX
BLOCK2 NUMERIC BYEAR BYEAR
MATCH2 CHAR LNAME LNAME .99 0.01
MATCH2 UNCERT FNAME FNAME .99 0.01 700
MATCH2 CHAR FINI FINI .96 0.01
MATCH2 PRORATED BMONTH BMONTH .99 0.01 1
MATCH2 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF2 26.0  20.0

BLOCK3 CHAR SDX SDX
BLOCK3 NUMERIC BYEAR BYEAR
MATCH3 UNCERT LNAME LNAME .99 0.01 700
MATCH3 UNCERT MNAME FNAME .99 0.01 700
MATCH3 CHAR MINI FINI .96 0.01
MATCH3 PRORATED BMONTH BMONTH .99 0.01 1
MATCH3 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF3 28.0  22.0

BLOCK4 CHAR SDX SDX
BLOCK4 CHAR FINI FINI
MATCH4 UNCERT LNAME LNAME .99 0.01 700
MATCH4 UNCERT FINI FINI .99 0.01 700
MATCH4 CHAR MINI MINI .96 0.01
MATCH4 CHAR SFIRST SFIRST .96 0.01
MATCH4 PRORATED BYEAR BYEAR .99 0.01 1
MATCH4 PRORATED BMONTH BMONTH .99 0.01 1
MATCH4 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF4 40.0  20.0
```

Figure 10   Matching Parameters for Male Records 1989-1992

```
BLOCK5 CHAR FINI FINI
BLOCK5 CHAR LNAME3 LNAME3
MATCH5 UNCERT LNAME LNAME .99 0.01 700
MATCH5 UNCERT FNAME FNAME .99 0.01 700
MATCH5 CHAR MINI MINI .96 0.01
MATCH5 PRORATED BMONTH BMONTH .99 0.01 1
MATCH5 PRORATED BDAY BDAY  .98 0.01 5
MATCH5 PRORATED BYEAR BYEAR .99 0.05 1
CUTOFF5 35.0  20.0

BLOCK6 CHAR SDX SDX
BLOCK6 CHAR BYEAR BYEAR
MATCH6 UNCERT LNAME LNAME .99 0.01 700
MATCH6 UNCERT FNAME FNAME .98 0.01 700
MATCH6 CHAR MINI MINI .96 0.01
MATCH6 CHAR SFIRST SFIRST .96 0.01
MATCH6 CHAR FFIRST FFIRST .96 0.01
MATCH6 CHAR MFIRST MFIRST .96 0.01
MATCH6 PRORATED BMONTH BMONTH .99 0.01 1
MATCH6 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF6 40.0  20.0

BLOCK7 CHAR LNAME3 LNAME3
BLOCK7 CHAR FINI FINI
MATCH7 UNCERT LNAME LNAME .99 0.01 700
MATCH7 UNCERT FNAME FNAME .99 0.01 700
MATCH7 UNCERT MNAME MNAME .71 0.01 700
MATCH7 CHAR SLAST SLAST .96 0.01
MATCH7 CHAR MINI MINI .96 0.01
MATCH7 PRORATED BYEAR BYEAR .98 0.01 1
MATCH7 PRORATED BMONTH BMONTH .99 0.01 1
MATCH7 PRORATED BDAY BDAY  .98 0.01 1
CUTOFF7 41.0  20.0
```

Figure 10 continued

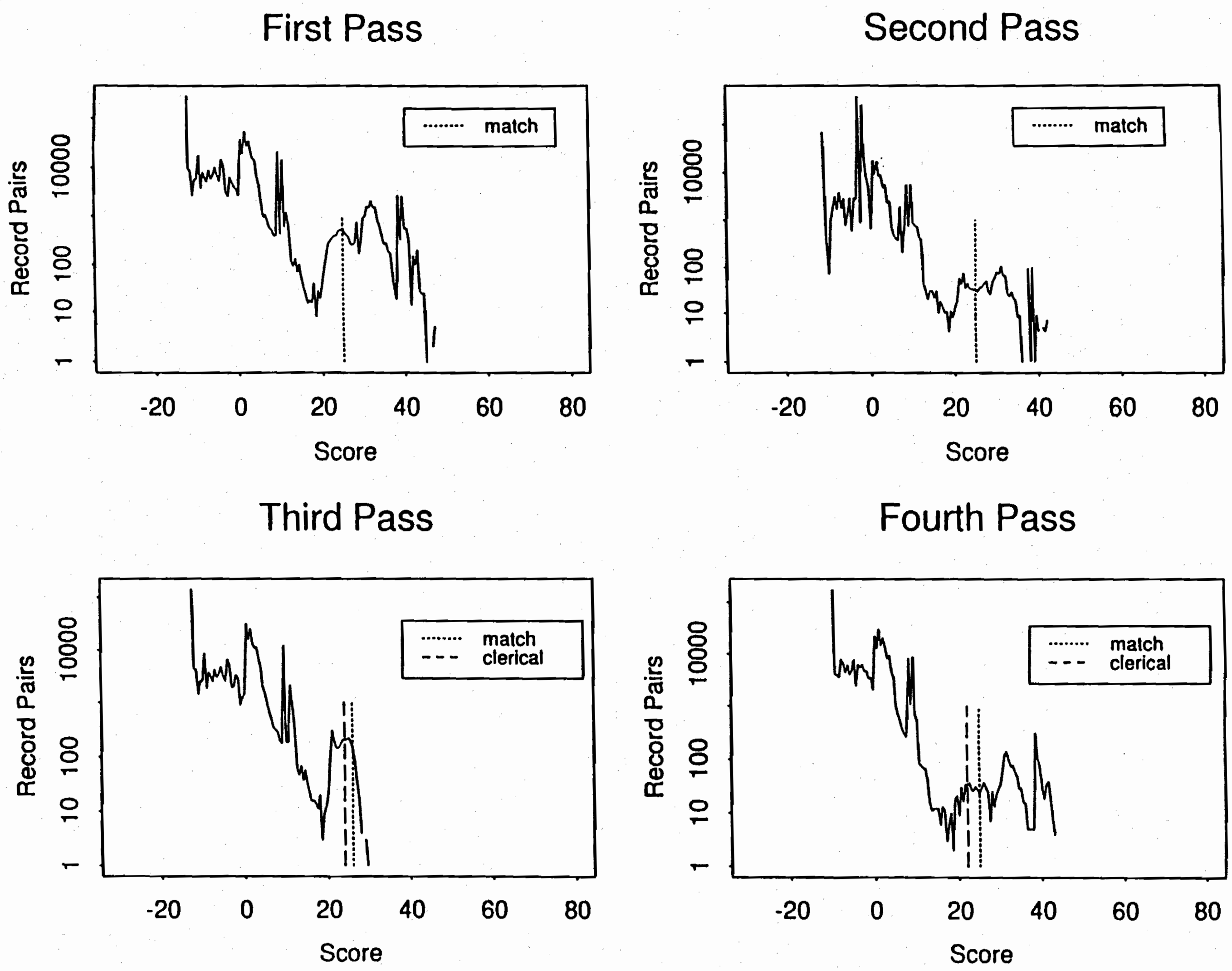


Figure 11 Female 1957-1979 Matching Histograms

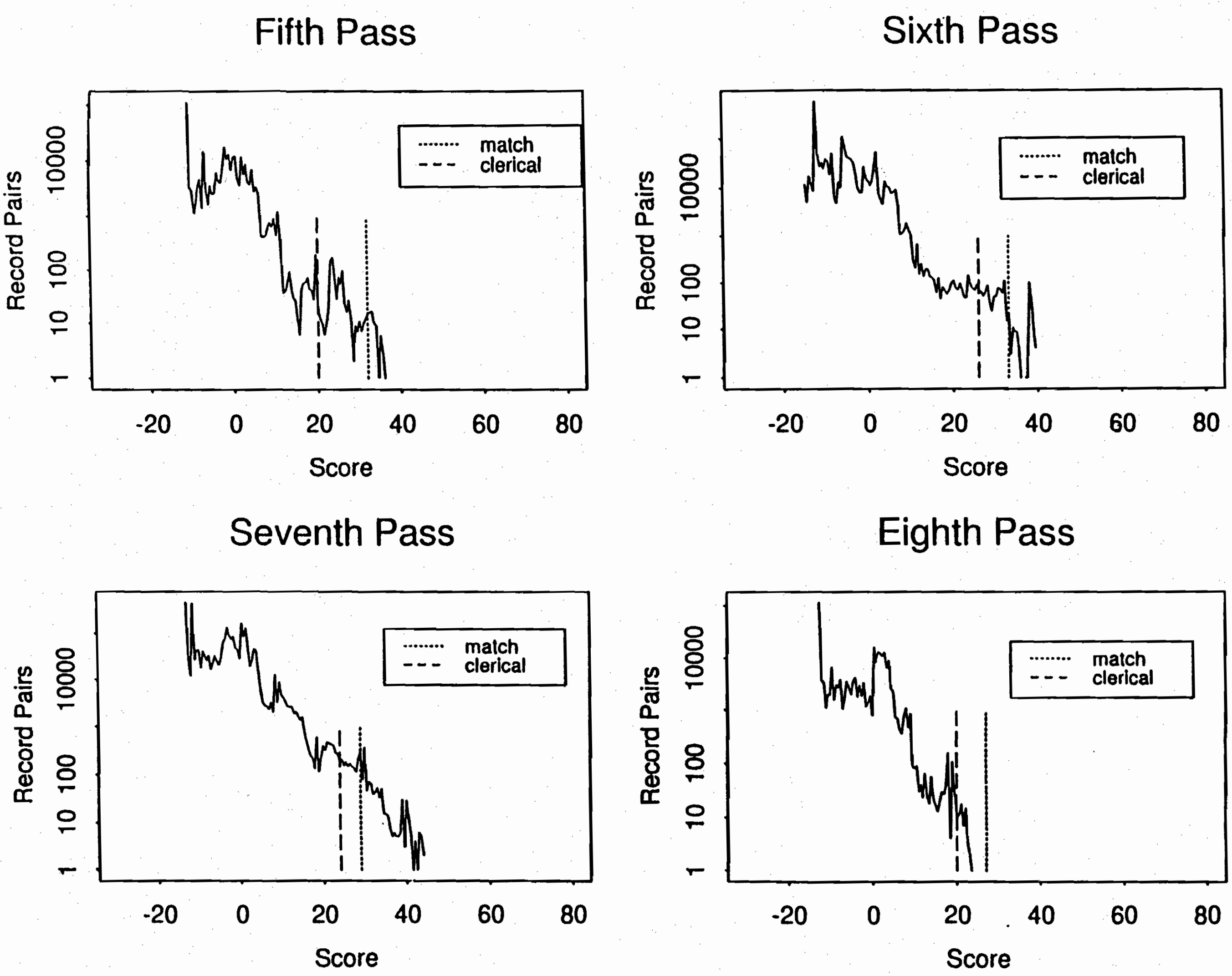


Figure 11 continued

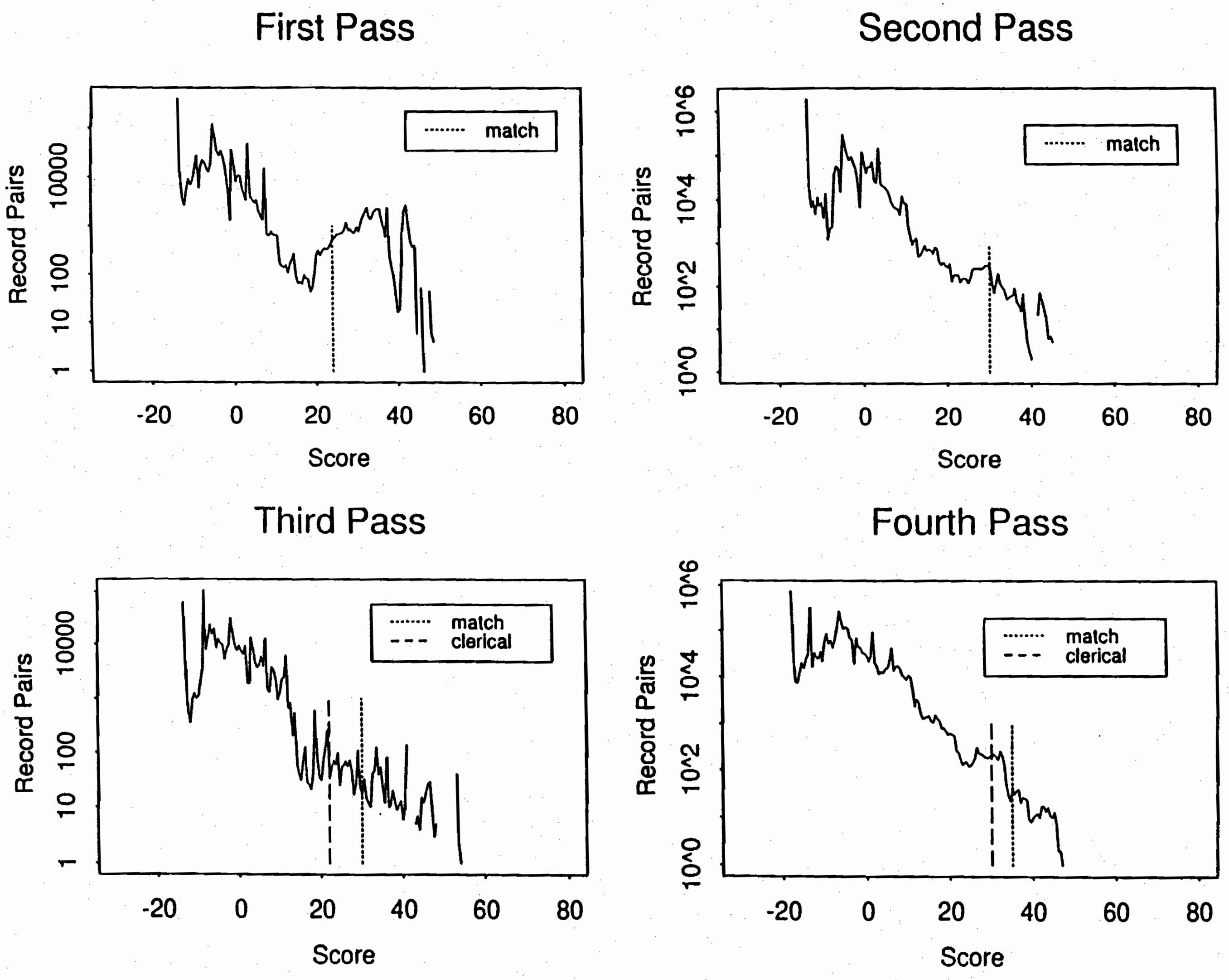
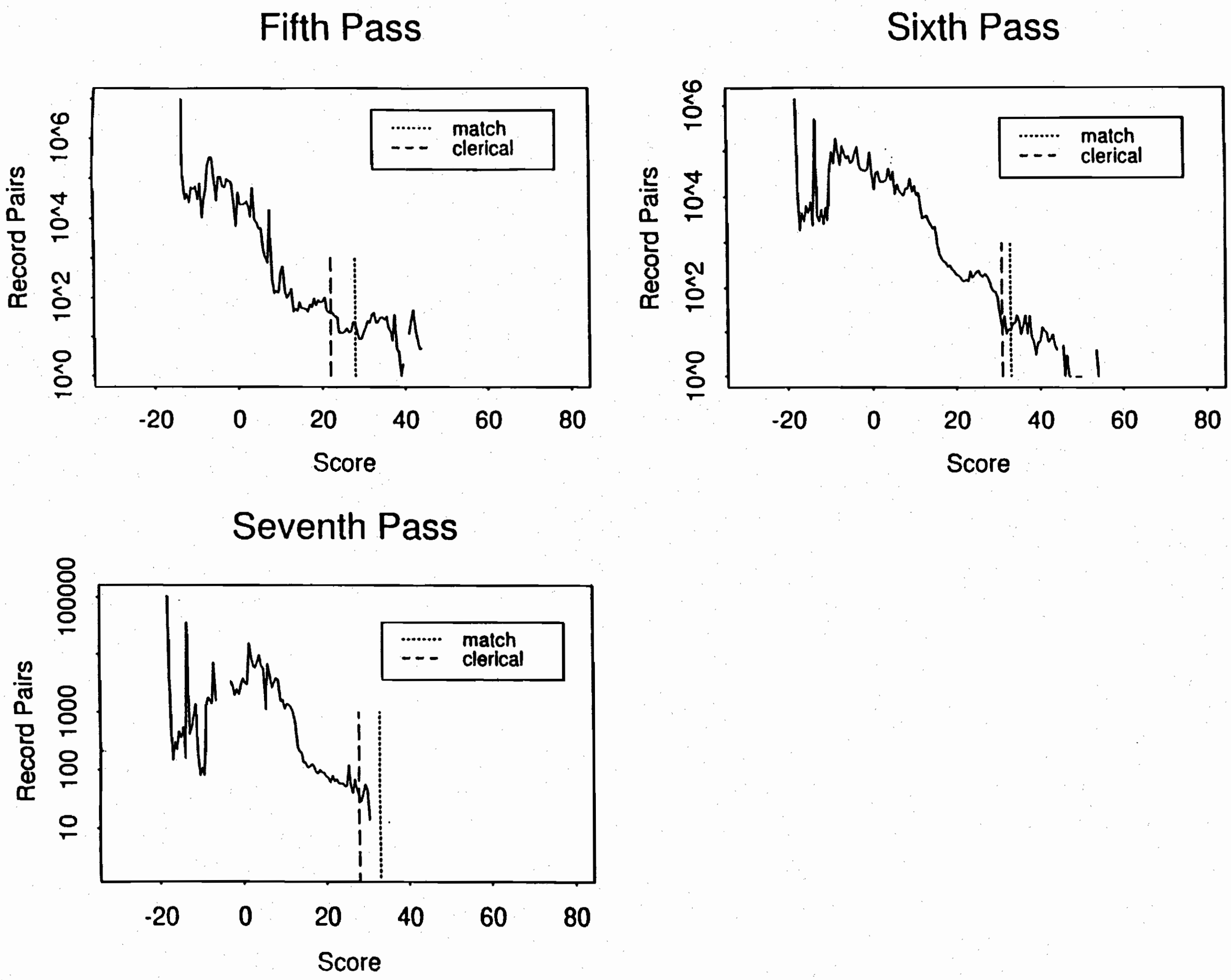


Figure 12 Male 1957-1979 Matching Histograms



Fifth Pass

Sixth Pass

Seventh Pass

Figure 12 continued

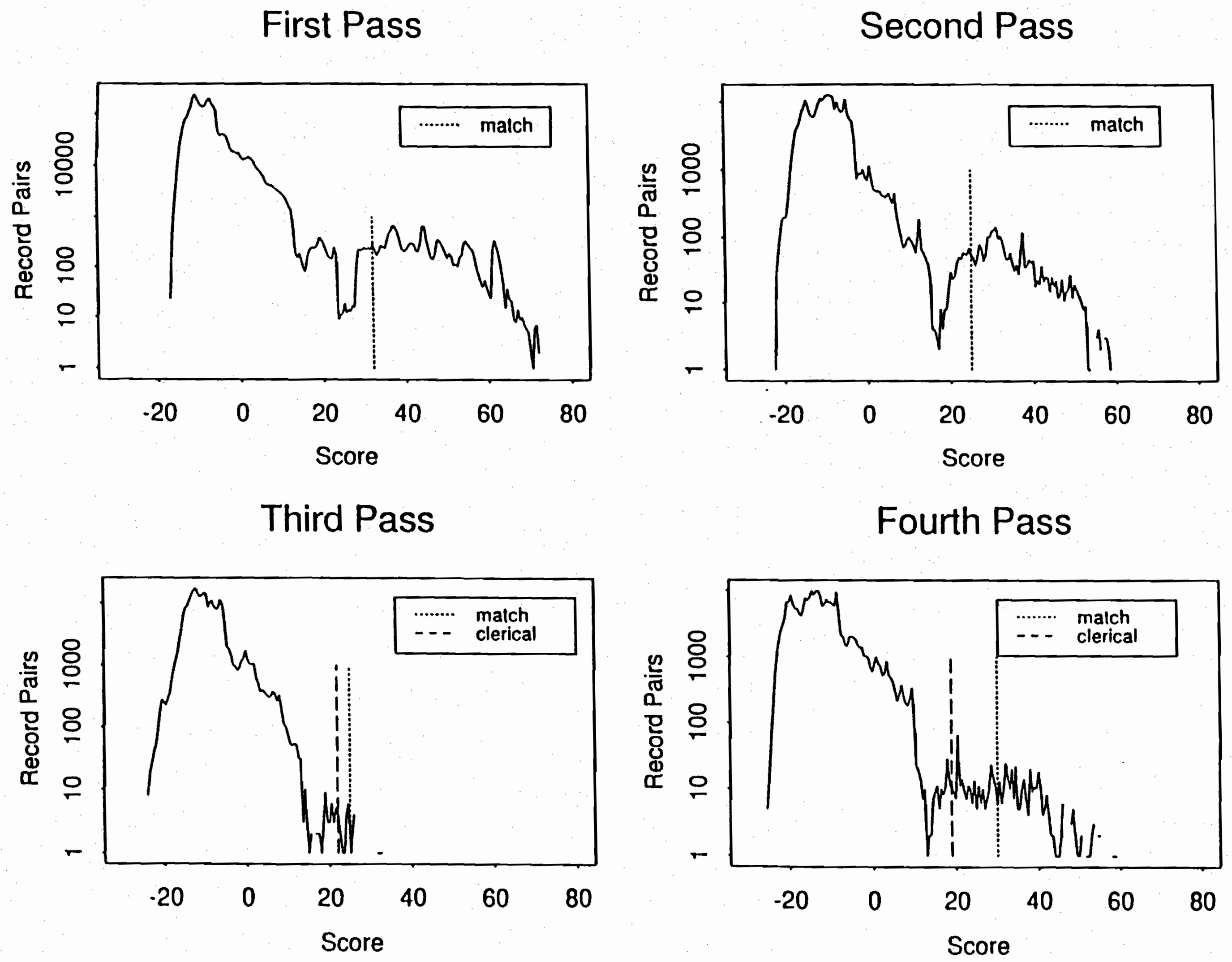
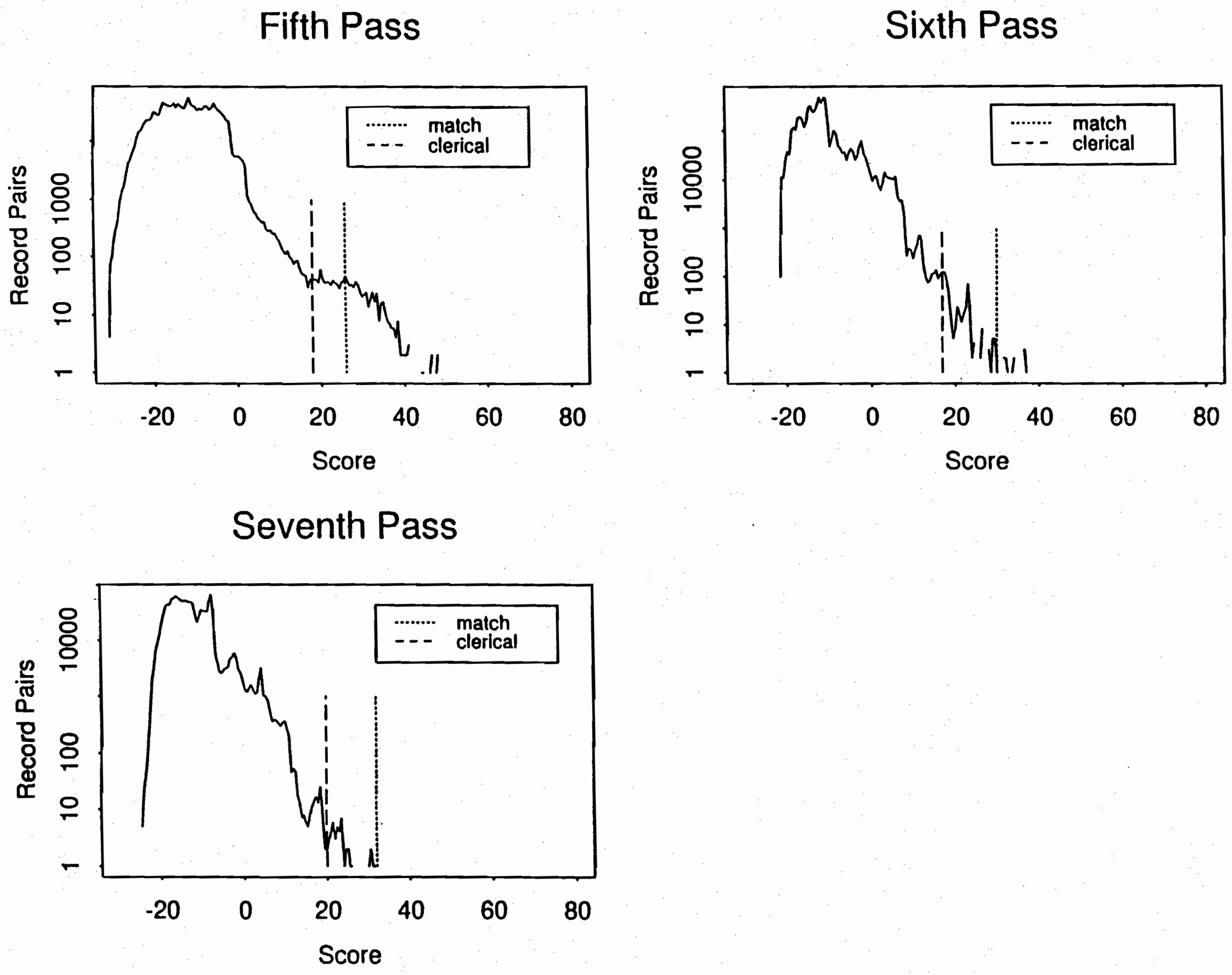


Figure 13 Female 1980-1988 Matching Histograms



Fifth Pass

Sixth Pass

Seventh Pass

Figure 13 continued

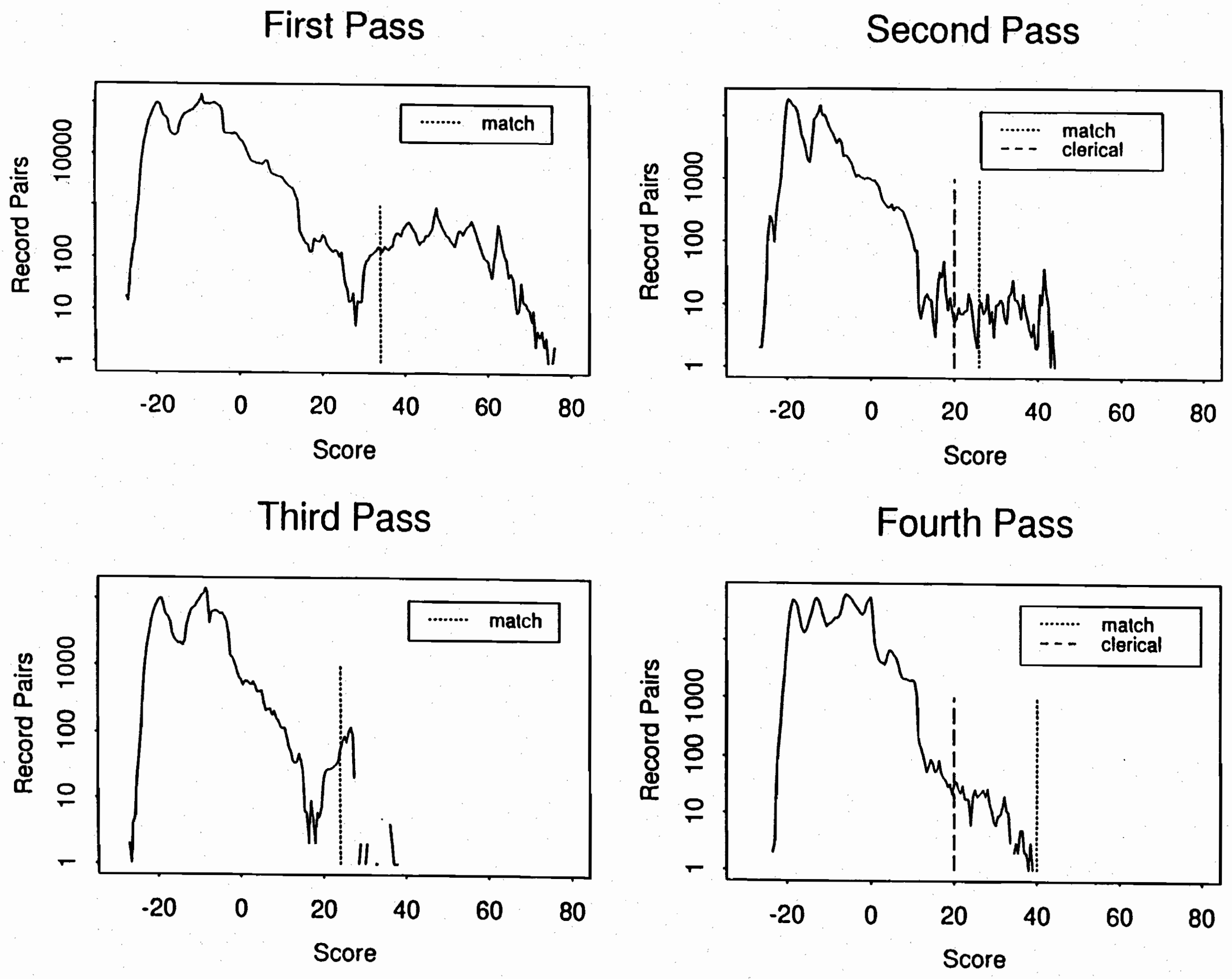


Figure 14 Male 1980-1988 Matching Histograms

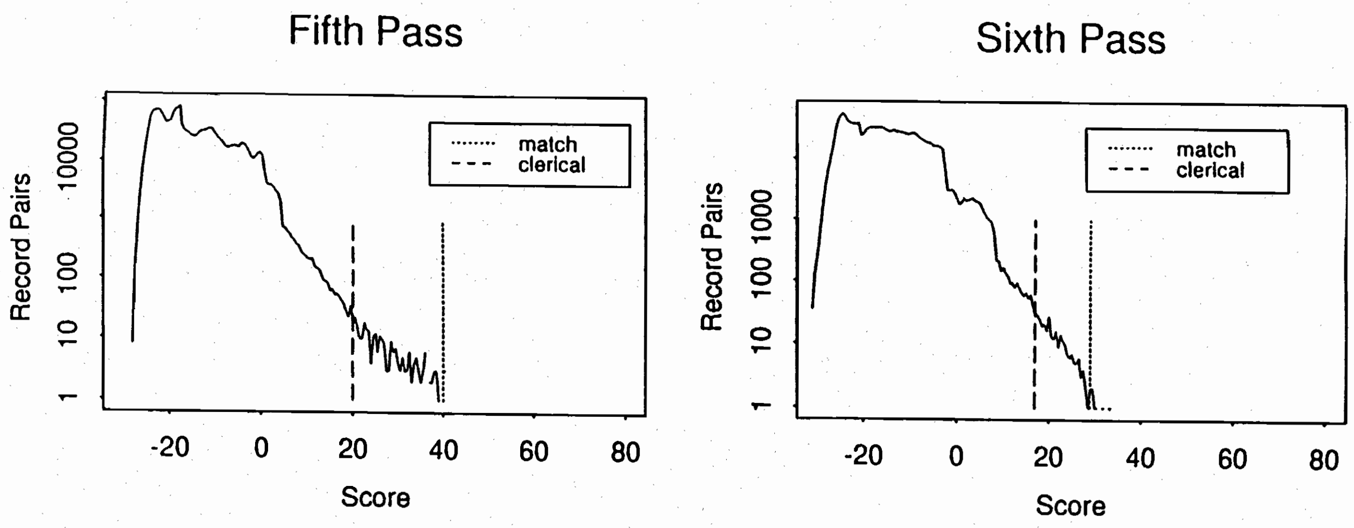


Figure 14 continued

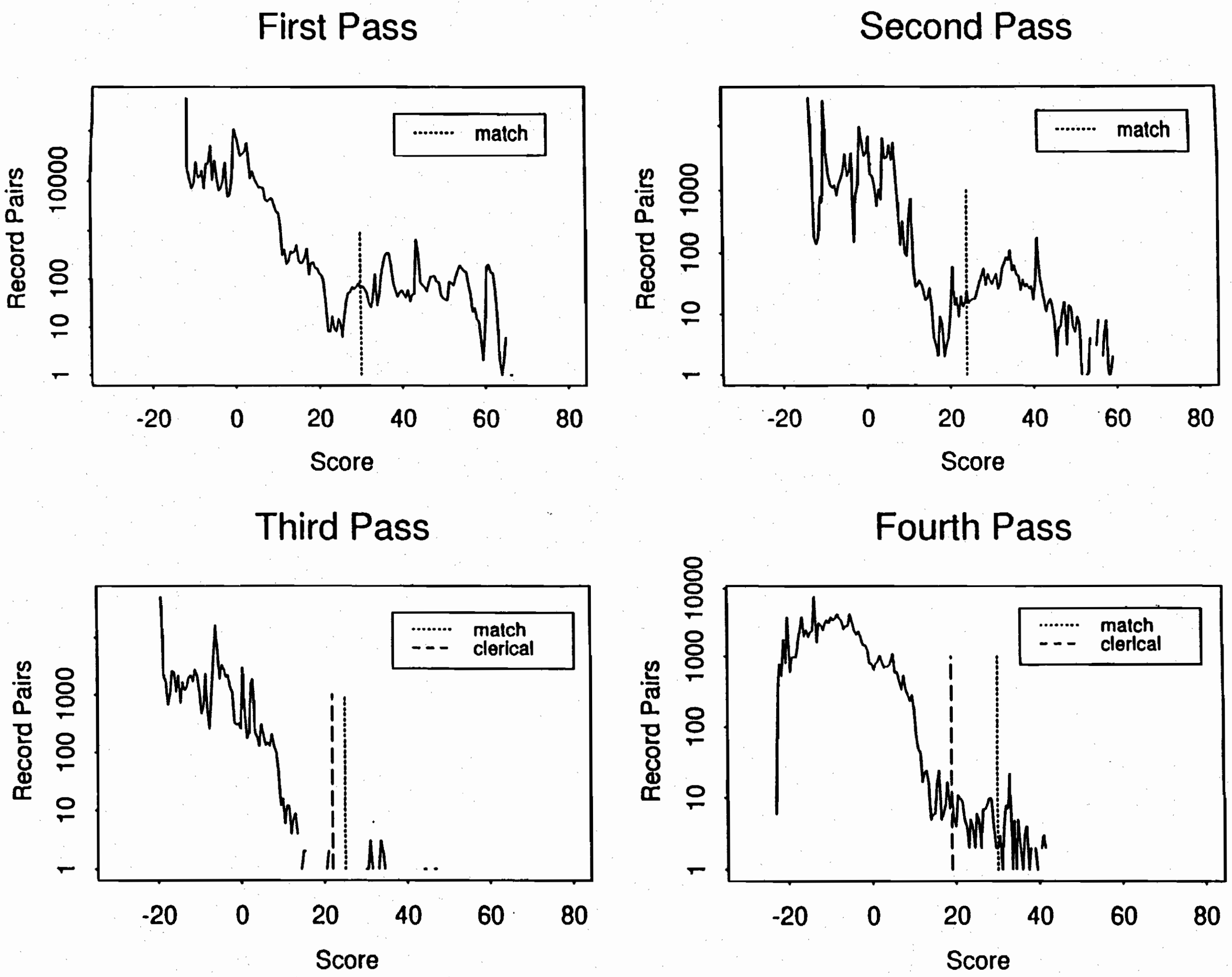


Figure 15 Female 1989-1992 Matching Histograms

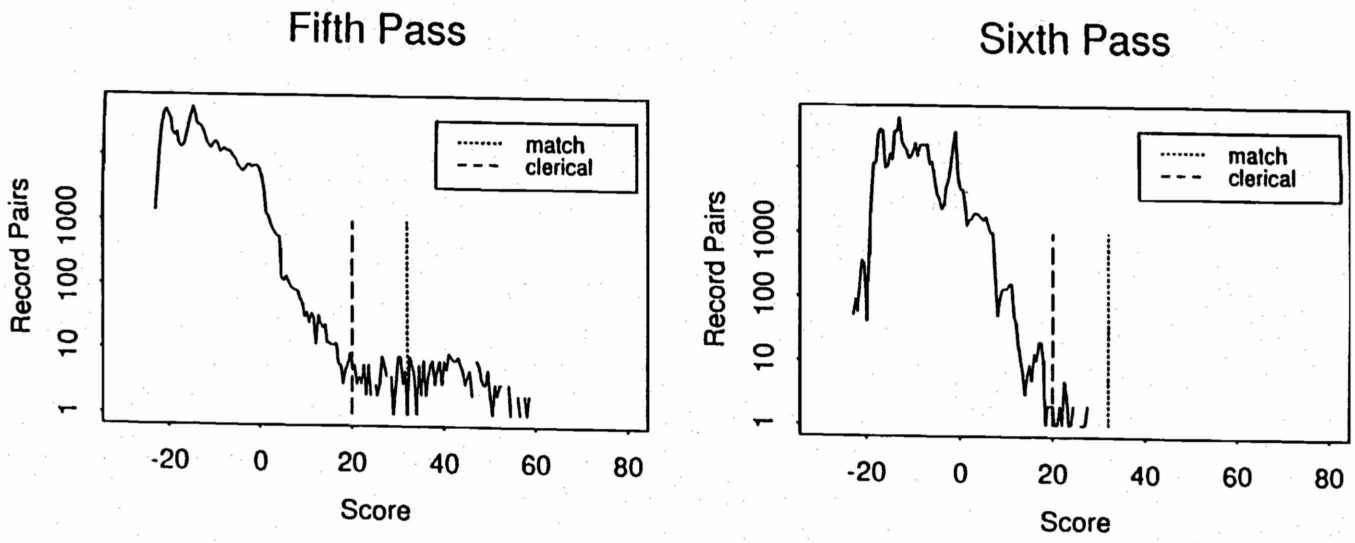
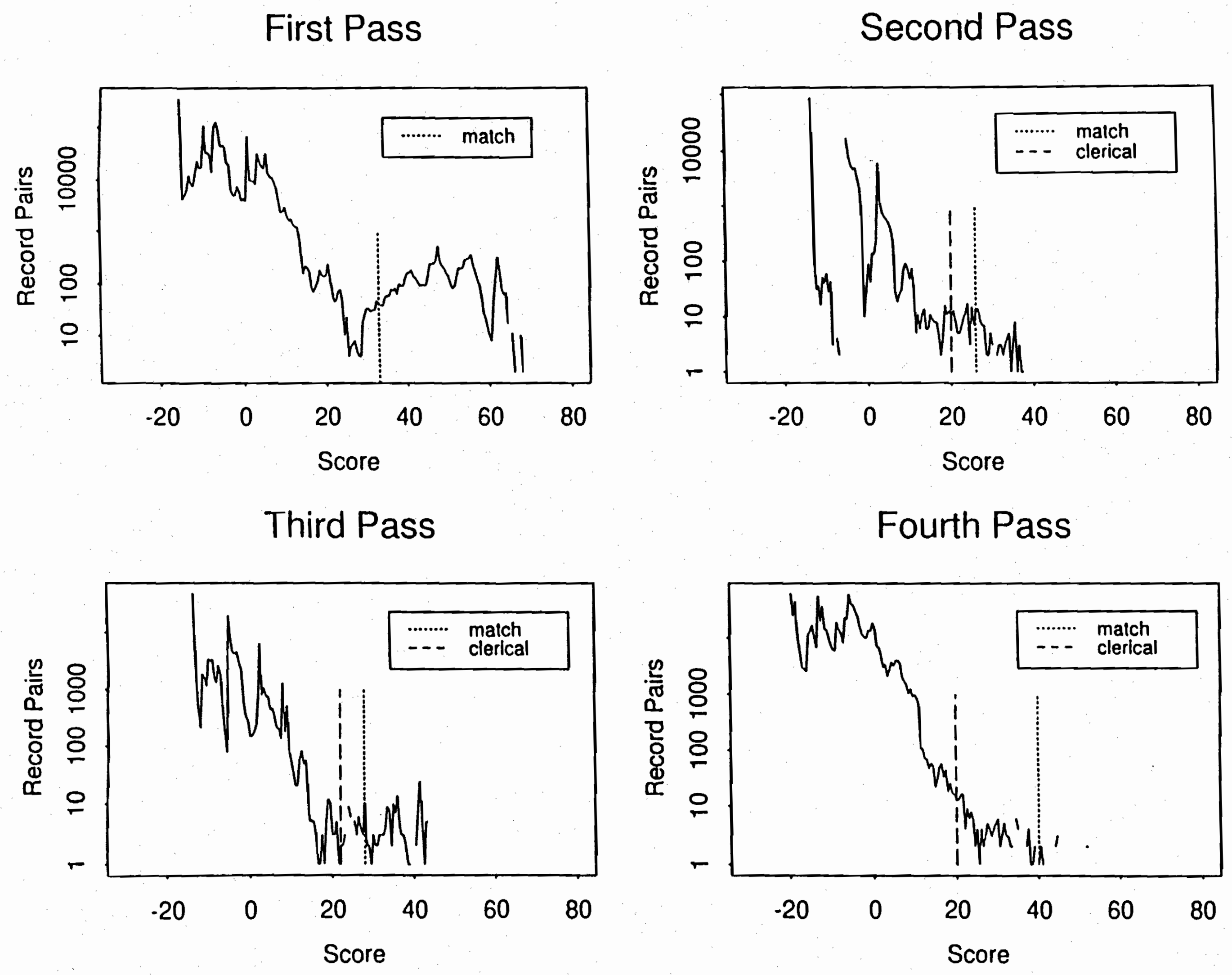


Figure 15 continued



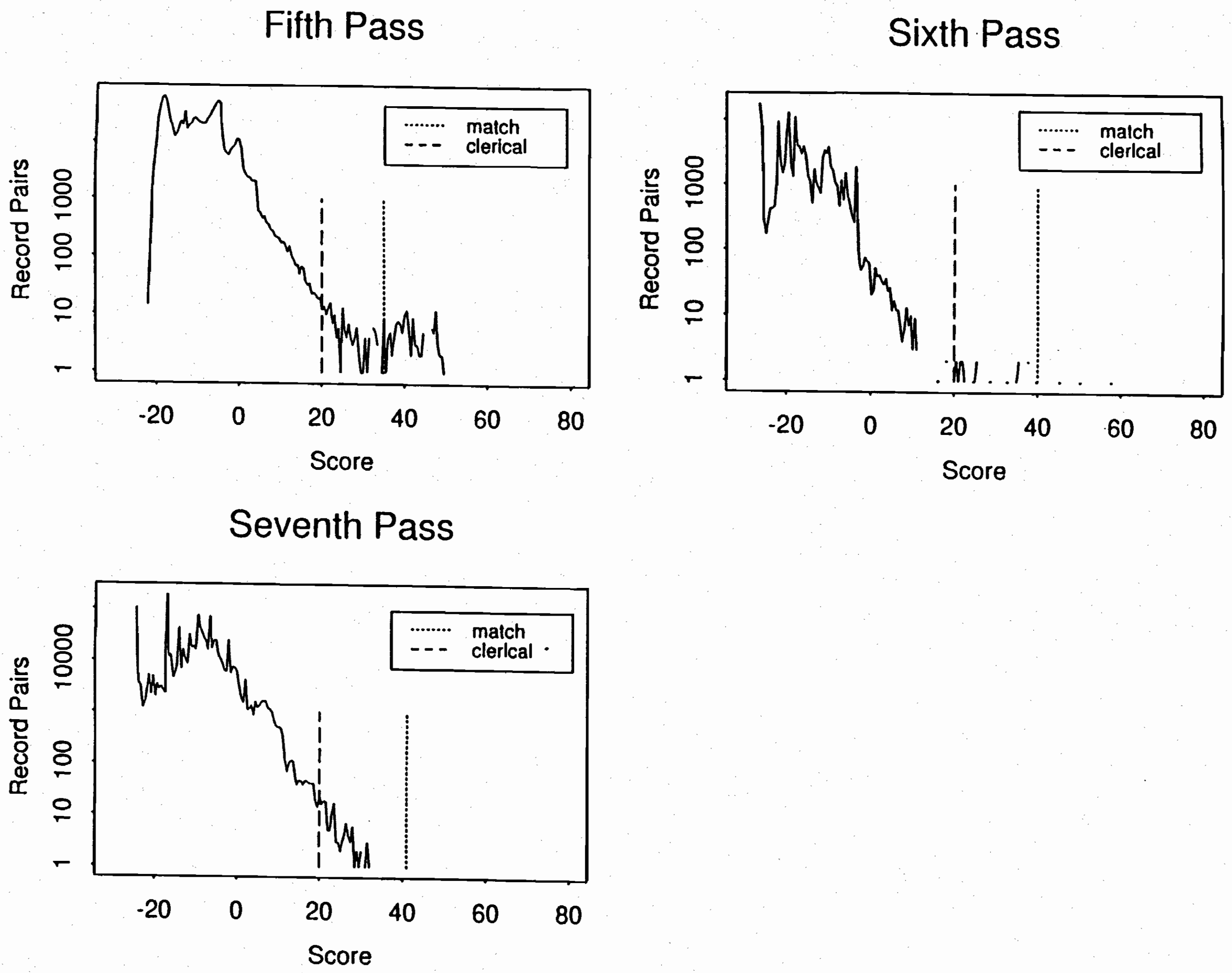
First Pass

Second Pass

Third Pass

Fourth Pass

Figure 16 Male 1989-1992 Matching Histograms


Fifth Pass

Sixth Pass

Seventh Pass

Figure 16 continued

vary by one.

The results of all of the matching runs for each death certificate set are shown in Table 4. There were 124,047 female death certificates and 158,715 male death certificates. There were 57,791 female death certificates and 68,114 male death certificates that linked to the genealogy records. The greatest percentage of females that linked came from the 1980-88 set, whereas the greatest percentage of male links came from the 1957-79 set. The lowest percentage for both groups came from the 1989-92 sets. This is most likely due to the lack of new information in the UPDB, since it was last updated in the early 1980s.

The percentage of links for the death certificates was much higher than the percentage for the cancer registry. Most of the difference can be attributed to the increased amount of information available with the death certificates. The death certificates contain additional fields such as the name of the individual's spouse,

Table 4 Death Certificate Linking Results

	Data Set	Death Certificates	Number Linked(%)
Females	1957-79	69,285	31,347 (45.2)
	1980-88	36,357	18,121 (49.8)
	1989-92	18,405	8,503 (46.2)
	Total	124,047	57,971 (46.7)
Males	1957-79	94,396	41,060 (43.5)
	1980-88	43,499	18,738 (43.1)
	1989-92	20,820	8,316 (39.9)
	Total	158,715	68,114 (42.9)
	Total All	282,762	126,085 (44.6)

mother and father. In the cancer records, 18% of the records have only a first and a last name. There is not a middle or name or maiden name. Only 6% of the death certificates do not have a middle initial or maiden name. The additional name fields contain powerful information that produce more links. Especially helpful is the maiden name for females, since every female record in the genealogy contains the maiden name.

Although there were more male links than female links, the percentage of male death certificates that linked was 42.9% compared to 46.7% for females. These results were different from the cancer linking where the linking percentage was substantially higher for males. Additional matching runs were tried for the males using all of the parameters such as spouse name and mother's name, but there was no significant increase in the number of males linked. Unlike the cancer records, there were similar amounts of information for each sex. Additional information such as father's first and last name and spouse's name increased the percentage of female links. In the female death records after 1980, only 1% of the records did not have a father's name. There was actually more information for the females, since the father's last name was the same for males.

There are 34,000 more male death certificates than female. The larger number of male death certificates is a result of the higher ratio of male to female births and the fact that more males were likely to migrate to the western states. There were 31,033 males and 24,907 females in the death certificates born outside of Utah, so there are 6,936 more males who migrated to Utah. Since the UPDB contains descendants of the Utah pioneers, it is likely that most of the recent immigrants to Utah are not in the UPDB.

The genealogy database was created in the mid-1970s, so anyone who was born after that would not be included in the database. Approximately 4% of the individuals in the 1989-92 set were born after 1980, so they would not be in the genealogy database and could not be linked. This was a likely cause of the lower linking percentages in the this set. There were 1,115 more males in this set of deaths, which would account for some of the reduced percentage of male links in these two sets.

Another measurement that helps to explain the lower percentage of males linked is the county of death. Several counties in Utah attract a larger number of immigrants and have a large number of individuals who would not be in the genealogy record, since it was associated with the LDS church. Many of the people who settled in Carbon county came to work in the mining industry. Most of the miners were not members of the LDS church, since mining was discouraged by the church. Salt Lake County is the largest county in the state. Since it is one of the major metropolitan areas in the mountain states and the state capital, it has attracted a diverse population. A large proportion of the population of Salt Lake county are not members of the LDS church. Weber county is another county that has a large non-LDS population that settled in the county to work in the railroad business and the military. A study of record linking with the census data from the 1880s has shown that individuals from these counties link to genealogical records at a much lower percentage when compared to other areas of the state (Mineau 1989). Census records from Utah and Cache counties, areas that were mostly rural communities settled by the pioneers, linked at a high percentage to the genealogical records. A summary of the linking rates for the death certificates of individuals who lived in these counties is shown in Table 5.

There are substantial differences in the linking percentages for the selected counties. Linking percentages in Cache and Utah counties are much higher than the other three counties. The percentages of linking for males and females in these counties are similar. In Carbon, Salt Lake, and Weber counties, the percentage of male links was always smaller. These counties would account for most of the differences in the overall linking percentages for males.

Table 5 Death Certificate Linking Results for Selected Counties

County	Sex	Number of Deaths	Number linked(%)
Cache	M	5,359	3,646 (68.0)
	F	4,650	3,206 (68.9)
Carbon	M	3,582	992 (27.7)
	F	2,159	695 (32.2)
Salt Lake	M	63,832	24,037 (37.7)
	F	53,106	22,058 (41.5)
Utah	M	16,434	9,487 (57.7)
	F	13,707	7,977 (58.2)
Weber	M	18,269	7,421 (40.6)
	F	14,427	6,368 (44.1)

The lower percentage of matches in the 1957-79 data sets was a result of missing data such as birth dates. As was mentioned previously, there were some years when birth year was not added to the death certificate record. In the records from 1957, 3,567 of the 5,929 death certificate records did not have any birth date information. Consequently only 1,611 (27.2%) of the records were linked. Table 6 and Figure 17 show this record linking results for all of the years. The lowest percentage of links are in 1959, 1960, and 1972 which correspond to having the highest number of death certificate records with no birth date information.

A birth year was calculated for the records that did not have a birth year using the age of death and death year. Depending on the accuracy of the age of death, the calculated birth year would be accurate to within one year. Approximately 200 additional links were added as a result. Those individuals with unusual names that generated higher scores made up most of the additional links, since the records with the calculated birth year were still missing a birth month and birth day.

Several reviews were done on the data to insure correct links. A report of all the links that received scores lower than 25 was reviewed manually, since there were some questions about the links in this range. The lowest score that could have been marked as a match was 20. Matches with a score above 25 were generally free of any ambiguity. All of the linked genealogy records that were twins were reviewed manually to make sure the correct twin was linked.

Table 6 Death Certificate Linking by Year of Death

Year of Death	Number of Deaths	Number without Birth Date	Number Linked	Percent Linked
1957	5,929	3,567	1,611	27.2
1958	6,010	414	2,759	45.9
1959	6,071	1,939	2,229	37.7
1960	6,240	1,822	2,298	36.8
1961	6,332	136	3,010	47.5
1962	6,509	59	3,198	49.1
1963	6,822	50	3,287	48.2
1964	6,594	118	3,184	48.3
1965	6,936	74	3,403	49.1
1966	7,087	21	3,497	49.3
1967	6,792	72	3,289	48.4
1968	7,149	70	3,341	46.7
1969	7,065	323	3,263	42.2
1970	7,351	19	3,514	47.8
1971	7,507	618	3,414	45.5
1972	7,560	2,804	2,495	33.0
1973	7,834	16	3,554	45.4
1974	7,703	8	3,563	46.3
1975	7,854	15	3,420	43.5
1976	7,810	14	3,468	44.4
1977	8,004	7	3,513	43.9
1978	8,257	8	3,519	42.6
1979	8,275	7	3,578	43.2
1980	8,479	2	3,767	44.4
1981	8,640	6	3,875	44.8
1982	8,860	6	3,927	44.3
1983	8,834	5	4,116	46.6
1984	9,319	5	4,255	45.7
1985	9,329	5	4,295	46.0
1986	9,330	10	4,139	44.4
1987	9,450	5	4,250	45.0
1988	9,594	4	4,235	44.1
1989	9,636	0	4,184	43.4
1990	9,511	0	4,113	43.2
1991	9,985	0	4,242	42.5
1992	10,114	0	4,280	42.3

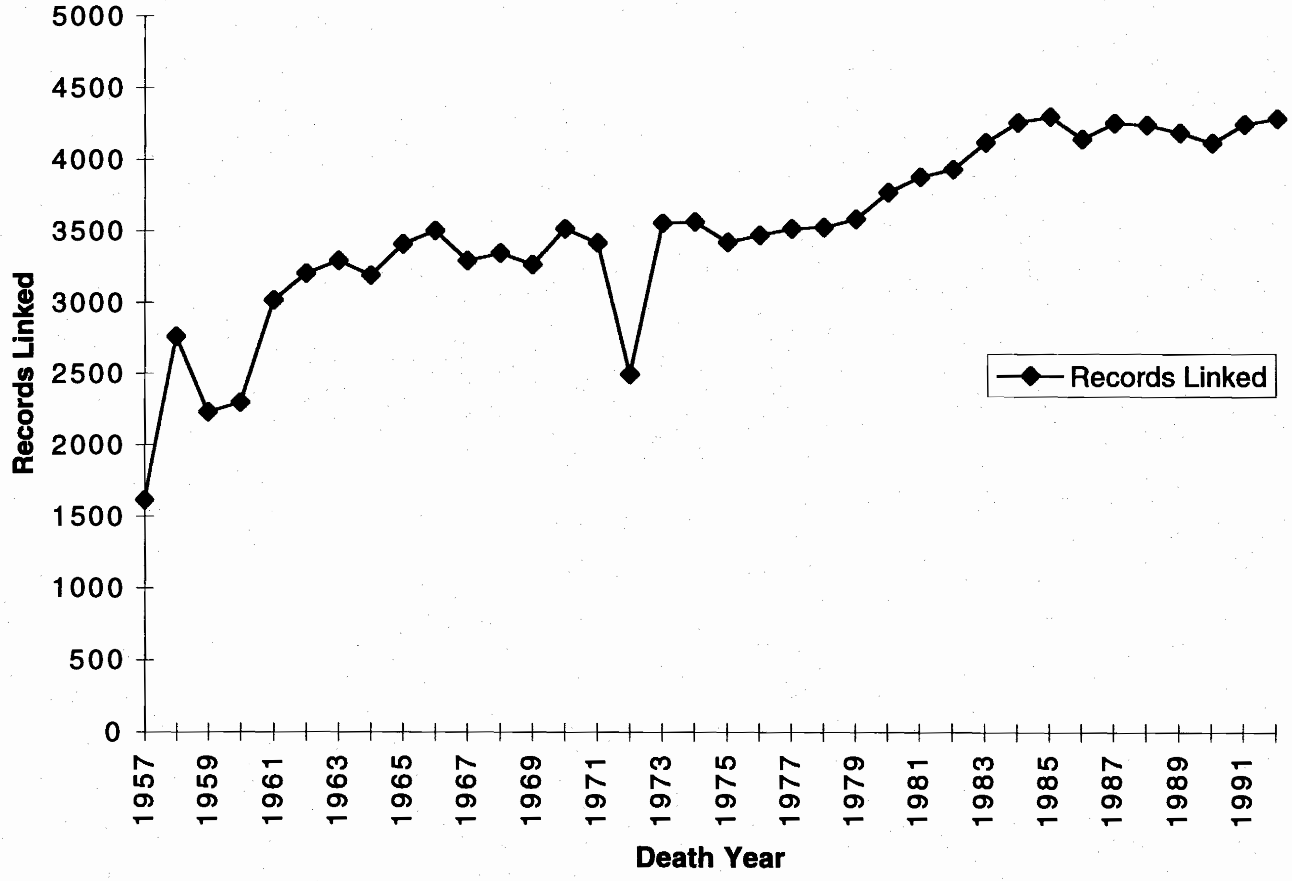


Figure 17 Death Certificate Linking by Year of Death

The second review was the calculation of a standard score for each of the records that linked. The score calculated in the linking process was not a standard score, since some of the matching parameters were not used in every pass to calculate the linking score. The standard score helps to determine the quality of the link, since it can be easily compared to the other links in the data set.

Since each of the three groups had different linking parameters and quality of data, a standard score was calculated separately for each group. The same number of parameters was used to generate a score for each match. The parameters that are listed in Table 3 were used for the standard score calculation. The last name, married name, and father's last name were all used for the females depending on which name was used to find the link. When the match was found because the middle name matched the first name, the standard score was calculated with the first name matching middle name, instead of first name matching first name and middle name matching middle name. This was done so that the scores would be generated by the same number of parameters for each match and the matches that came from different combinations of names would not have lower scores. The lowest 1000 scores from each group were reviewed manually and links that were determined to be incorrect were deleted from the matched set.

The average scores were 43.64, 45.87 and 58.97 for the 1957-79, 1980-88, and 1989-92 sets respectively. The 1980-88 set did not have as much data as the other two sets, but the quality of the data was better than the oldest set. The 1989-92 set had the most complete data, which is reflected in its scores.

The last step of the death certificate linking was to create a file that had all of the information needed for future studies. The information in this file included identification numbers, death dates, birth dates, ICD codes for cause of death, ICD revision numbers, and the age of death. Also included were the standard score and the data set the death certificate came from.

There are a number of duplicate individual records in the genealogy database. Many of these records do not link to ancestors. For the final file, a program was run to take the linked duplicates and select the record that was linked to parents in the genealogy. The record that linked to a parent was used since it is essential in constructing family pedigree information for use in future genetic studies.

CHAPTER 4

GENEALOGICAL INDEX OF FAMILIALITY AND RELATIVE RISK

The linked death certificates were grouped by cause of death. The cause of death groups were analyzed by the genealogical index of familiality and first-degree relative risk to explore the familial relationships for each cause of death. This was done to show the usefulness of the linked death certificates and to produce a comprehensive examination of the familial aggregation of the causes of death.

The medical, environmental, and inherited aspects of each cause of death were researched. The results of the analyses of the linked death records were discussed with respect to the findings from this research.

4.1 Cause of Death Classification

The first step in the genetic epidemiological analysis of the death certificate records was the classification of the causes of death. Closely related causes of death were grouped together because they could be a result of the same genetic predisposition or environmental condition. For example, hypertensive heart disease and hypertensive renal disease were both classified as hypertension. Each group contained a minimum of 100 cases. This limit was established since smaller numbers would not produce meaningful results in the epidemiological analysis. Some causes of death that were not analyzed, because there were less than 100 cases, included thyroid cancer, pharynx cancer, epilepsy, and hepatitis. The groups and the associated causes of death, along with the average age of death, are listed in Table 7. Those groups that have more than one cause of death combined into a single group are shown in Table 8.

There were 92,774 primary and 7,748 secondary distinct causes of death put into 61 groups. Diabetes, hypertension, and pneumonia account for 7,432 of the secondary causes of death. Approximately 75% of the linked death records were put into a group based on the cause of death. The remaining death records had a cause of death that did not have a large enough group or had a cause of death which was not to studied. For example, there are more than 4,000 deaths from causes such as falls, poisonings, accidental shootings, fires, industrial accidents, aircraft accidents, and drowning that were not studied. Deaths from motor vehicle accidents were studied for an example of a cause of death that would not be genetically predisposed, but the other accidental deaths were not studied.

There were three different ICD code revisions used during the time period of the death certificates. This made selecting specific death records tedious, since the ICD codes changed between each revision. The ICD codes were recoded to

Table 7 Cause of Death Groups

Cause of Death	Number	Average Age	Standard Deviation
Heart Disease	20,480	76.39	11.54
Myocardial Infarction	13,542	73.65	10.93
Stroke	8,211	80.39	9.58
Pneumonia	6,357	80.14	12.94
Diabetes	6,013	74.48	11.97
Hypertension	4,954	77.72	10.67
Motor Vehicle	2,514	48.10	23.86
Prostate Cancer	2,481	77.47	8.56
Colon Cancer	2,246	71.98	12.06
Breast Cancer	2,203	65.51	13.90
Lung Cancer	2,120	69.11	10.28
Congestive Heart Failure	2,053	83.19	9.27
Conduction Disorders	1,782	78.58	11.69
Suicide	1,441	49.79	17.51
Chronic Airway Obstruction	1,278	75.57	8.48
Emphysema	1,180	70.92	9.27
Pancreatic Cancer	1,171	70.91	11.44
Stomach Cancer	989	70.72	12.79
Lymphoma	969	68.67	14.55
Aneurysm	930	73.78	11.26
Heart Valve Disorders	870	63.19	14.29
Ovarian Cancer	738	66.06	12.22
Pulmonary Embolism	716	71.49	13.36
Cirrhosis	705	63.49	13.10
Ulcer	680	72.74	13.35
Renal Failure	675	78.32	12.62
Senility without Psychosis	653	86.49	6.76
Brain Cancer	646	57.36	18.00
Alzheimer's Disease	564	81.15	8.13
Myeloma	553	69.86	11.12
Parkinson's Disease	505	77.78	7.56
Bladder Cancer	501	75.43	10.07
Endocarditis	495	76.81	12.30
Intestinal Obstruction	483	76.43	13.59
Nephritis	482	65.86	19.05
Biliary Tract and Gallbladder Disorders	462	74.94	12.79
Myeloid Leukemia	461	65.65	16.74
Rectal Cancer	457	71.72	11.99
Cardiomyopathy	452	72.10	15.22
Kidney Cancer	407	67.28	14.57
Uterine Cancer	416	70.09	12.14
Alcohol Related	399	60.72	13.72
Circulatory Disorder	394	73.18	15.18
Melanoma	382	62.20	16.58
Bronchitis	374	70.82	17.36
Senility with Psychosis	346	83.55	9.04
Asthma	301	70.01	15.13

Table 7(continued)

Cause of Death	Number	Average Age	Standard Deviation
Lymphoid Leukemia	292	70.30	18.32
Liver Cancer	292	68.42	13.35
Mouth Cancers	277	69.88	11.92
Congenital Anomalies of Circulatory System	268	20.10	26.91
Influenza	260	77.15	17.14
Gallbladder Cancer	243	72.84	11.41
Esophageal Cancer	228	67.92	11.53
Obesity	226	65.60	13.39
Hodgkin's Disease	204	55.51	19.13
Motor Neuron Diseases	202	69.05	10.25
Multiple Sclerosis	186	56.61	12.90
Diverticulosis	179	78.43	10.49
Connective Tissue Cancer	163	64.31	17.66
Cervical Cancer	166	62.98	15.52

Table 8 Cause of Death Groups with Multiple Causes

Cause of Death Group	Includes
Mouth Cancers	malignant neoplasm of lip, tongue, salivary glands, floor of mouth, larynx, oropharynx, gum
Lymphoma	lymphosarcoma, reticulosarcoma, Burkitt's lymphoma
Alcohol Related	alcoholic psychosis, acute alcoholic intoxication, alcohol abuse
Motor Neuron Disease	amyotrophic lateral sclerosis, progressive muscular atrophy, bulbar palsy
Heart Valve Disorders	diseases and disorders of mitral, aortic, tricuspid valves
Hypertension	hypertensive heart and renal disease
Heart Disease	ischemic heart disease, angina pectoris, coronary atherosclerosis
Conduction Disorders	heart conduction disorders, cardiac dysrhythmia
Stroke	subarachnoid, intracerebral, intracranial hemorrhages, cerebral arteries occlusion, cerebral embolism, cerebrovascular disease
Aneurysm	aortic aneurysm, cerebral aneurysms
Circulatory Disorders	peripheral vascular disease, arterial embolism and thrombosis, other disorders of arteries and arterioles
Pneumonia	viral, pneumoccal, bacterial pneumonia
Ulcer	gastric, duodenal, peptic ulcers
Biliary Tract and Gallbladder Disorders	cholelithiasis, cholecystis, other biliary tract disorders

eliminate this problem. Each ICD code was assigned a number that corresponded to the 61 groups. If a ICD code was not assigned a group, it was set to zero.

4.2 Genealogical Index of Familiality Methods

The GIF has been used with the UPDB records to study cancer and heart disease (Williams 1978; Hill 1980; Cannon 1982). The familiality of cancer in Utah was studied again with updated cancer records and the genealogical index in 1994 (Cannon-Albright 1994). In this study, the linked cancer records were used to measure the familial clustering of cancer. The methods of that study were used for a similar study with the death certificate data.

The GIF was developed to measure the degree of family clustering in the UPDB. The GIF measures the degree of relationship between all possible pairs of individuals in a group by using the Malécot coefficient of kinship to quantify the degree of relatedness of two individuals (Malécot 1948). The coefficient of kinship for each pair is defined as the probability that randomly selected homologous genes from the two individuals are identical by descent from a common ancestor. The calculation of the kinship is made by counting total paths of descent. Each path contributes an exponent of 1/2 to the total kinship. The value of the exponent is equal to the number of individuals along the path. For example, the kinship of two siblings would be the sum of $1/2^3$ and $1/2^3$, since both siblings would be related through each parent and there are three individuals along the path between the siblings. The kinship of half-siblings would be $1/2^3$, since they would only be related through one parent. The mean of the coefficients of kinship for all pairs of cases is multiplied by 10^5 to give the single measure of familiality called the GIF.

The kinship for the linked death certificates was calculated for all of the cases in each disease group. It was also calculated for subgroups of each group and combinations of disease groups. The subgroups were the male cases, female cases, and approximately the youngest third of the group based on age of death. The kinship for each group was compared to the kinship of a set of randomly selected, matched controls. This is necessary to produce a meaningful comparison of the observed familiality that takes into account the sample size and the demographic characteristics of each individual in the disease group. The kinship by itself is meaningless, since it has no dimensions or units. It gains meaning when compared to a control group.

The controls were matched to the cases by birth year, sex, and birthplace. The criteria for matching by birthplace were those individuals born in Utah and those born outside of Utah. The controls were chosen at random from the UPDB according to these matching criteria. The distribution of the kinship and the mean of the controls kinship coefficient varies randomly, depending on which controls were chosen, so control groups were selected 100 times and kinship calculations repeated. The repeated calculations of the controls give an empirical distribution for the control GIF. The mean of the coefficients of kinship from the death cases is compared with the mean of these 100 calculations. Under the hypothesis of no familial aggregation for the disease, the kinship case GIF is a random observation from the distribution of the control GIF values.

This empirical method allows a significance test for excess familiality of any of the linked death record groups. This significance can be calculated by

$$Z = [(\text{case GIF - mean control GIF})/\text{control SD}]$$

when compared with a standard normal distribution. The Z score calculated by this equation is a measure of where the case GIF lies in the distribution of the control GIF values. It can be used to produce a one-sided p-value that shows the probability that the difference between the case GIF and the mean control GIF value is not a result of chance.

The standard deviation (SD) of the controls is affected by the sample size. Sets of controls that are small will have a large range of GIF values and therefore have a larger standard deviation than a control set with a larger number of individuals. The fact that the Z score is affected by the sample size is important since a group of closely related cases in a small sample can produce a large value for the GIF that may not be indicative of the true familiality of a disease. However causes of death with large sample sizes can produce small p-values with very small differences between the case GIF and control GIF. The interpretation of the p-value for the GIF should be made with the size of the sample considered.

The use of the birth year, sex, and birthplace as control selection factors has been studied previously with Utah Cancer Registry records (Cannon 1982). This study examined a number of methods for control selection, including the method used here. The other methods included the requirement that controls be alive when the cancer registry began or that a death certificate be available for a control and that the death occurred after the cancer registry was begun. The study determined that all of the methods gave similar results consistent with the random variation inherent with the control selection.

Requiring certain fields insured that the control sets and the cause of death sets would both have complete UPDB records and thus be closely matched. There are records in the UPDB that are missing vital information such as birth year. These incomplete records in the UPDB are often not linked into a family, so they were not included as possible controls since they would not contribute anything to the control GIF. If these records had been used they may have caused an underestimation of the GIF in the control sets. The use of birth year required that a control has a birth year and thus insured a more complete record. Record linking also required a birth year, as there were no links made if the death certificate record did not have a birth year.

4.3 Genealogical Index of Familiality Results

The GIF data for all of the causes of death groups are shown in Tables 9 and 10. Table 9 is ordered by the GIF value and Table 10 is ordered by p-value.

Table 9 GIF for All Linked Death Certificate Records (ordered by GIF)

Cause of Death	N	Death Cases GIF	Control Cases GIF	P-Value*
Multiple Sclerosis	**186**	**6.86**	**3.11**	**0.0002**
Kidney Cancer	**407**	**5.44**	**3.02**	**0.0000**
Congenital Anomalies	**268**	**5.10**	**2.58**	**0.0000**
Hodgkin's Disease	**204**	**4.95**	**2.83**	**0.0186**
Influenza	260	4.37	2.79	0.0563
Alcohol Related	399	4.34	2.86	0.0005
Motor Neuron Diseases	202	4.16	2.75	0.0647
Myeloma	**553**	**4.15**	**2.98**	**0.0034**
Mouth Cancers	**277**	**4.12**	**2.94**	**0.0728**
Aneurysm	**930**	**4.02**	**2.90**	**0.0000**
Gallbladder Cancer	**243**	**4.00**	**3.03**	**0.1428**
Asthma	301	3.97	2.87	0.0640
Chronic Airway Obstruction	1,278	3.95	2.92	0.0000
Lymphoid leukemia	292	3.95	2.75	0.0569
Parkinson's Disease	505	3.91	2.85	0.0021
Prostate Cancer	2,481	3.88	2.90	0.0000
Ovarian Cancer	738	3.87	2.84	0.0003
Myeloid Leukemia	461	3.86	2.84	0.0058
Cardiomyopathy	452	3.75	2.85	0.0204
Diverticulosis	179	3.73	2.91	0.2665
Emphysema	1,180	3.72	2.97	0.0019
Suicide	1,441	3.71	2.81	0.0000
Senility Without Psychosis	653	3.69	2.74	0.0265
Lymphoma	969	3.67	2.78	0.0000
Diabetes	6,014	3.64	2.89	0.0000
Pulmonary Embolism	716	3.57	2.88	0.0125
Stomach Cancer	976	3.56	2.81	0.0014
Nephritis	482	3.56	2.90	0.0785
Circulatory Disorders	394	3.54	2.93	0.1765
Cervical Cancer	166	3.53	2.89	0.3091
Brain Cancer	646	3.50	2.78	0.0112
Obesity	226	3.34	2.92	0.3363
Cirrhosis	705	3.33	2.91	0.1042
Alzheimer's Disease	564	3.31	2.95	0.1907
Lung Cancer	2,120	3.28	2.86	0.0003
Congestive Heart Fail	2,053	3.28	2.95	0.0173
Heart Valve Disorders	870	3.21	2.91	0.1302
Ulcer	680	3.21	2.94	0.2156
Endocarditis	495	3.17	2.86	0.2365
Stroke	8,211	3.16	2.95	0.0003
Colon Cancer	2,246	3.14	2.87	0.0116
Motor Vehicle Accident	2,514	3.11	2.66	0.0000

Table 9 (continued)

Cause of Death	N	Death Cases GIF	Control Cases GIF	P-Value*
Myocardial Infarction	13,543	3.07	2.91	0.0001
Liver Cancer	292	3.06	2.89	0.4071
Rectal Cancer	457	3.04	2.86	0.3625
Hypertension	4,954	3.04	2.87	0.0219
Breast Cancer	2,203	3.03	2.81	0.0450
Melanoma	382	2.99	2.82	0.3703
Pneumonia	6,357	2.97	2.89	0.1062
Biliary,Gallbladder Disorders	462	2.92	2.90	0.4844
Pancreatic Cancer	1,171	2.86	2.81	0.4016
Uterine Cancer	416	2.86	2.97	0.5760
Heart Disease	20,480	2.81	2.91	0.9977
Heart Conduction Disorders	1,782	2.80	2.94	0.8550
Renal Failure	675	2.80	2.89	0.5968
Bladder Cancer	501	2.62	2.95	0.7359
Esophageal Cancer	228	2.56	2.87	0.6275
Connective Tissue Cancer	163	2.49	2.86	0.6204
Senility with Psychosis	346	2.17	3.01	0.8606
Intestinal Obstruction	483	2.14	2.90	0.9457
Bronchitis	374	2.07	2.86	0.9260

*P-value for hypothesis that death cases GIF > control GIF

Table 10 GIF for All Linked Death Certificate Records
(ordered by p-value)

		Death	Control	
		Cases	Cases	
Cause of Death	N	GIF	GIF	P-Value*
Kidney Cancer	**407**	**5.44**	**3.02**	**0.0000**
Congenital Anomalies	**268**	**5.10**	**2.58**	**0.0000**
Aneurysm	**930**	**4.02**	**2.90**	**0.0000**
Chronic Airway Obstruction	**1,278**	**3.95**	**2.92**	**0.0000**
Suicide	**1,441**	**3.71**	**2.81**	**0.0000**
Prostate Cancer	**2,481**	**3.88**	**2.90**	**0.0000**
Lymphoma	**969**	**3.67**	**2.78**	**0.0000**
Diabetes	**6,014**	**3.64**	**2.89**	**0.0000**
Motor Vehicle Accident	**2,514**	**3.11**	**2.66**	**0.0000**
Myocardial Infarction	**13,543**	**3.07**	**2.91**	**0.0001**
Multiple Sclerosis	186	6.86	3.11	0.0002
Ovarian Cancer	738	3.87	2.84	0.0003
Lung Cancer	2,120	3.28	2.86	0.0003
Stroke	**8,211**	**3.16**	**2.95**	**0.0003**
Alcohol Related	399	4.34	2.86	0.0005
Stomach Cancer	976	3.56	2.81	0.0014
Emphysema	1,180	3.72	2.97	0.0019
Parkinson's Disease	505	3.91	2.85	0.0021
Myeloma	553	4.15	2.98	0.0034
Myeloid Leukemia	461	3.86	2.84	0.0058
Brain Cancer	646	3.50	2.78	0.0112
Colon Cancer	2,246	3.14	2.87	0.0116
Pulmonary Embolism	716	3.57	2.88	0.0125
Congestive Heart Fail	2,053	3.28	2.95	0.0173
Hodgkin's Disease	204	4.95	2.83	0.0186
Cardiomyopathy	452	3.75	2.85	0.0204
Hypertension	4,954	3.04	2.87	0.0219
Senility Without Psychosis	653	3.69	2.74	0.0265
Breast Cancer	2,203	3.03	2.81	0.0450
Influenza	260	4.37	2.79	0.0563
Lymphoid leukemia	292	3.95	2.75	0.0569
Asthma	301	3.97	2.87	0.0640
Motor Neuron Diseases	202	4.16	2.75	0.0647
Mouth Cancers	277	4.12	2.94	0.0728
Nephritis	482	3.56	2.90	0.0785
Cirrhosis	705	3.33	2.91	0.1042
Pneumonia	6,357	2.97	2.89	0.1062
Heart Valve Disorders	870	3.21	2.91	0.1302
Gallbladder Cancer	243	4.00	3.03	0.1428
Circulatory Disorders	394	3.54	2.93	0.1765
Alzheimer's Disease	564	3.31	2.95	0.1907
Ulcer	680	3.21	2.94	0.2156
Endocarditis	495	3.17	2.86	0.2365

Table 10 (continued)

Cause of Death	N	Death Cases GIF	Control Cases GIF	P-Value*
Diverticulosis	179	3.73	2.91	0.2665
Cervical Cancer	166	3.53	2.89	0.3091
Obesity	226	3.34	2.92	0.3363
Rectal Cancer	457	3.04	2.86	0.3625
Melanoma	382	2.99	2.82	0.3703
Pancreatic Cancer	1,171	2.86	2.81	0.4016
Liver Cancer	292	3.06	2.89	0.4071
Biliary,Gallbladder Disorders	462	2.92	2.90	0.4844
Uterine Cancer	416	2.86	2.97	0.5760
Renal Failure	675	2.80	2.89	0.5968
Connective Tissue Cancer	163	2.49	2.86	0.6204
Esophageal Cancer	228	2.56	2.87	0.6275
Bladder Cancer	501	2.62	2.95	0.7359
Heart Conduction Disorders	1,782	2.80	2.94	0.8550
Senility with Psychosis	346	2.17	3.01	0.8606
Bronchitis	374	2.07	2.86	0.9260
Intestinal Obstruction	483	2.14	2.90	0.9457
Heart Disease	20,480	2.81	2.91	0.9977

*P-value for hypothesis that death cases GIF > control GIF

Where the p-values are equal in Table 10, the cause of death with the highest GIF is listed first. The different ordering schemes are provided to give two different views of the GIF results. The values in the table include the number of cases, case GIF, control mean GIF, and the p-value. Causes of death that are discussed in the text are in bold type.

The highest values come from the multiple sclerosis, kidney cancer, congenital anomalies, and hodgkin's disease groups. The lowest scores come from bronchitis, intestinal obstruction, and senility with psychosis. Some of the problems in interpreting the GIF are a result of small sample sizes. Since it is possible for one or two sets of relatives to inflate the kinship in a small sample, the mean of the kinship coefficient would be high, since the high kinship value would be divided over a small number of cases. For the causes of death with the 12 highest GIF values, only the kidney cancer, aneurysm, and myeloma groups have more than 400 cases. However, most of the causes of death with the highest GIF values do have small p-values. The causes of death that do have large p-values and large GIF values such as gallbladder cancer and mouth cancers have sample sizes less than 300.

When the GIF results are ordered by p-values, there are a number of causes of death that move to the top of the list. These include chronic airway obstruction, aneurysm, suicide, prostate cancer, lymphoma, and diabetes. Kidney cancer and congenital anomalies remain at the top of the list. Other causes of death at the top of the list have large numbers of cases such as motor vehicle accidents, myocardial infarction, and stroke. There is some evidence of familiality for these causes of death, but the low p-values are partly due to the large sample sizes. The low p-values from the causes of death with large sample sizes are an indication that the GIF score is more reliable than the score from a cause of death with a small sample size. The table that is ordered by p-value is likely the most accurate ranking of the GIF scores. A more detailed examination of the results for each cause of death is given in section 4.8.

Approximately one-third of the youngest cases for each cause of death group was selected by age of death. There are some diseases, such as coronary heart disease, whose risk of the disease in a family increases greatly if there are family members who develop the disease at an early age (Jorde 1995). A high GIF value for the youngest group may signify genetic predispositions. A GIF value was calculated for each of these youngest groups. The results are in Table 11. The causes of death are ordered by the GIF value. Motor neuron diseases and connective tissue cancer have the highest GIF values, but they both have a very small sample size, so the reliability of the GIF values is suspect. The causes of death with high GIF values and sample sizes greater than 200 include emphysema, aneurysm, ovarian cancer, chronic airway obstruction, and myeloma. Other causes of death with significant GIF values are Alzheimer's disease, suicide, prostate cancer, and diabetes.

Table 12 lists the causes of death where there was a large difference between the GIF for the younger cases and all of the cases. Large differences

Table 11 GIF for Youngest Third of Linked Death Certificate Records

Cause of Death (Oldest age of death in years)	N	Death Cases GIF	Control Cases GIF	P-Value*
Motor Neuron Diseases(65)	**68**	**14.75**	**2.70**	**0.0000**
Connective Tissue Cancer(60)	**53**	**10.38**	**2.75**	**0.0099**
Emphysema(65)	**328**	**8.00**	**2.91**	**0.0000**
Asthma(70)	123	7.57	2.95	0.0004
Kidney Cancer(65)	146	7.21	2.90	0.0009
Aneurysm(70)	**326**	**6.85**	**2.96**	**0.0000**
Multiple Sclerosis(55)	81	6.25	2.64	0.0257
Diverticulosis(75)	64	6.20	2.98	0.1710
Ovarian Cancer(60)	**215**	**6.19**	**2.87**	**0.0000**
Liver Cancer(65)	106	6.11	2.75	0.0125
Cardiomyopathy(67)	155	5.85	3.21	0.0152
Parkinson's Disease(75)	185	5.75	2.86	0.0076
Mouth Cancers(70)	138	5.52	2.65	0.0142
Hodgkin's Disease(55)	91	5.29	2.98	0.0844
Chronic Airway(70)	**349**	**5.22**	**3.02**	**0.0004**
Myeloma(70)	**289**	**5.20**	**2.93**	**0.0006**
Alzheimer's Disease(80)	**262**	**4.89**	**3.04**	**0.0064**
Suicide(40)	**465**	**4.75**	**2.78**	**0.0000**
Prostate Cancer(73)	**753**	**4.61**	**2.91**	**0.0000**
Stomach Cancer(67)	343	4.45	2.95	0.0031
Pulmonary Embolism(70)	294	4.38	2.83	0.0086
Diabetes(70)	**1,803**	**4.36**	**2.88**	**0.0000**
Congenital Anomalies(2)	129	4.25	2.46	0.1084
Nephritis(65)	187	4.23	2.84	0.1082
Ulcer(70)	258	4.03	3.01	0.1322
Bladder Cancer(70)	146	4.03	3.00	0.2334
Lymphoma(65)	325	3.92	2.85	0.0461
Congestive Heart Failure(79)	590	3.91	2.87	0.0011
Motor Vehicle Accident(40)	994	3.61	2.68	0.0000
Gallbladder Cancer(70)	87	3.60	2.74	0.3338
Lung Cancer(64)	669	3.47	2.94	0.0567
Myocardial Infarction(69)	4,447	3.44	2.90	0.0000
Colon Cancer(66)	657	3.43	2.88	0.0492
Senility Without Psych(83)	201	3.40	2.95	0.3635
Brain Cancer(60)	317	3.34	2.77	0.1902
Cirrhosis(60)	279	3.34	2.88	0.2504
Uterine Cancer(70)	195	3.32	3.16	0.4494
Myeloid Leukemia(65)	190	3.31	2.59	0.2225
Pneumonia(75)	1,586	3.27	2.87	0.0073
Hypertension(73)	1,486	3.26	2.93	0.0467
Endocarditis(73)	150	3.24	2.71	0.3367
Influenza(75)	92	3.17	2.72	0.3930
Breast Cancer(60)	821	3.15	2.81	0.1178
Renal Failure(75)	211	3.13	2.88	0.4031
Heart Disease(71)	6,124	3.08	2.93	0.0074
Stroke(75)	2,097	3.08	2.96	0.1754
Pancreatic Cancer(65)	352	3.08	2.78	0.2844

Table 11 (continued)

Cause of Death (Oldest age of death)	N	Death Cases GIF	Control Cases GIF	P-Value*
Lymphoid Leukemia(70)	112	3.06	2.79	0.4258
Heart Valve Disorders(60)	363	2.83	2.96	0.5480
Cervical Cancer(56)	55	2.70	3.19	0.5476
Heart Conduction(75)	597	2.56	2.89	0.8141
Biliary Tract, Gallbladder Disorders(70)	139	2.51	2.96	0.6172
Bronchitis(70)	162	2.00	2.68	0.7508
Melanoma(60)	168	1.71	2.80	0.8301
Rectal Cancer(70)	190	1.57	3.14	0.8954
Alcohol Related(60)	189	1.46	3.25	0.9623
Intestinal Obstruction(74)	161	1.31	2.80	0.8916
Senility with Psychosis(80)	101	1.08	2.54	0.7910
Circulatory Disorders(70)	136	1.03	2.72	0.8984
Obesity(60)	76	0.84	3.02	0.8399
Esophageal Cancer(63)	77	0.73	3.17	0.8300

*P-value for hypothesis that death cases GIF > control GIF

Table 12 GIF Values for All Death Certificates Compared to Youngest Third of Death Certificates (ordered by youngest cases GIF)

Cause of Death	All Cases GIF	All Cases P-Value*	Youngest Cases GIF	Youngest Cases P-value*
Motor Neuron Diseases	**4.16**	**0.0647**	**14.75**	**0.0000**
Connective Tissue Cancer	**2.49**	**0.6204**	**10.38**	**0.0099**
Emphysema	**3.72**	**0.0019**	**8.00**	**0.0000**
Asthma	**3.97**	**0.0640**	**7.57**	**0.0004**
Kidney Cancer	5.44	0.0000	7.21	0.0009
Aneurysm	4.02	0.0000	6.85	0.0000
Diverticulosis	**3.73**	**0.2665**	**6.20**	**0.1710**
Ovarian Cancer	**3.87**	**0.0003**	**6.19**	**0.0000**
Liver Cancer	3.06	0.4071	6.11	0.0125
Cardiomyopathy	3.75	0.0204	5.85	0.0152
Parkinson's Disease	3.91	0.0021	5.75	0.0076
Mouth Cancers	4.12	0.0728	5.52	0.0142
Chronic Airway Obstruct.	3.95	0.0000	5.22	0.0004
Hodgkin's Disease	4.95	0.0186	5.22	0.0004
Myeloma	4.15	0.0034	5.20	0.0006
Alzheimer's Disease	3.31	0.1907	4.89	0.0064
Suicide	3.71	0.0000	4.75	0.0000
Prostate Cancer	3.88	0.0000	4.61	0.0000
Stomach Cancer	3.56	0.0014	4.45	0.0031
Pulmonary Embolism	3.57	0.0125	4.38	0.0086
Diabetes	3.64	0.0000	4.36	0.0000
Nephritis	3.56	0.0785	4.23	0.1082
Ulcer	3.21	0.2156	4.03	0.1322
Bladder Cancer	2.62	0.7359	4.03	0.2334
Lymphoma	3.67	0.0000	3.92	0.0461
Congestive Heart Failure	3.28	0.0173	3.91	0.0011
Motor Vehicle Accident	3.11	0.0000	3.61	0.0000
Lung Cancer	3.28	0.0003	3.47	0.0567
Myocardial Infarction	3.07	0.0001	3.44	0.0000
Colon Cancer	3.14	0.0116	3.43	0.0492
Uterine Cancer	2.86	0.5760	3.32	0.4494
Hypertension	3.04	0.0219	3.26	0.0467
Pneumonia	2.97	0.1062	3.27	0.0073
Breast Cancer	3.03	0.0450	3.15	0.1178
Renal Failure	2.80	0.5968	3.13	0.4031
Heart Disease	2.81	0.9977	3.08	0.0074

*P-value for hypothesis that death cases GIF > control GIF

were observed in motor neuron diseases, connective tissue cancer, emphysema, asthma, diverticulosis, and ovarian cancer.

The GIF for males and females from each cause of death was calculated to look for any sex-related differences in familiality. The GIF values for the males are listed in Table 13 and the female values are in Table 14. Table 15 lists those causes of death where a substantial difference between the males and females was seen. The highest familiality for males was seen for influenza, congenital anomalies, gallbladder cancer, alcohol related, and myeloid leukemia. The highest values for the females were seen with Hodgkin's disease, multiple sclerosis, kidney cancer, and aneurysm. A substantial difference in GIF values was seen for suicide where there were five times as many male deaths and the GIF for males was substantially higher. Also interesting was lung cancer where there were four times as many male deaths, but the GIF for females was higher.

An analysis of the GIF for combined sets was done to look for any interactions between the disease groups. It was done by combining two cause of death groups. If there were any duplicates in the combined file, the duplicate was removed and the GIF calculated. A duplicate could occur where an individual had both causes of death listed on their death certificate. The results are summarized in Table 16. Only a portion of the combined scores are shown in the table. Some of the scores were chosen to be in the table, because the two diseases have been shown to be medically or genetically linked such as breast and ovarian cancer. The most interesting combinations are those where the

Table 13 GIF for Male Linked Death Certificate Records

Cause of Death	N	Death Cases GIF	Control Cases GIF	P-Value*
Influenza	**117**	**7.57**	**3.09**	**0.0131**
Congenital Anomalies	**134**	**7.47**	**2.58**	**0.0000**
Gallbladder Cancer	**94**	**6.34**	**2.52**	**0.0233**
Connective Tissue Cancer	82	5.97	3.00	0.1275
Kidney Cancer	**246**	**5.90**	**2.83**	**0.0000**
Diverticulosis	62	5.78	3.25	0.2174
Alcohol Related	**324**	**5.44**	**2.94**	**0.0000**
Mouth Cancers	208	5.41	2.92	0.0072
Myeloid Leukemia	**242**	**4.98**	**2.72**	**0.0011**
Circulatory Disorders	190	4.90	2.73	0.0303
Liver Cancer	140	4.83	2.66	0.0688
Myeloma	314	4.53	2.95	0.0110
Biliary Tract, Gallbladder	220	4.17	2.90	0.1203
Nephritis	286	4.10	2.80	0.0493
Obesity	78	4.16	2.93	0.3090
Suicide	**1,180**	**4.02**	**2.83**	**0.0000**
Endocarditis	259	3.90	2.87	0.1120
Prostate Cancer	2,481	3.88	2.90	0.0000
Asthma	142	3.87	2.97	0.2243
Pulmonary Embolism	377	3.85	2.77	0.0167
Emphysema	1,062	3.84	2.83	0.0004
Diabetes	2,559	3.81	2.86	0.0000
Parkinson's Disease	325	3.77	2.86	0.0796
Lymphoid leukemia	182	3.73	2.99	0.2700
Senility Without Psych	253	3.71	2.68	0.1445
Aneurysm	701	3.51	2.97	0.0604
Cirrhosis	397	3.47	2.93	0.1589
Stomach Cancer	575	3.44	2.81	0.0366
Stroke	3,368	3.41	2.96	0.0000
Chronic Airway Obstruction	981	3.41	2.94	0.0508
Lymphoma	523	3.29	2.73	0.0945
Lung Cancer	1,674	3.28	2.87	0.0040
Hypertension	1,876	3.25	2.91	0.0111
Heart Valve Disorders	395	3.25	2.84	0.2231
Motor Vehicle Accident	1,676	3.23	2.72	0.0001
Brain Cancer	356	3.17	2.75	0.1983
Colon Cancer	1,112	3.15	2.88	0.0997
Cardiomyopathy	245	3.14	2.99	0.4298
Myocardial Infarction	8,711	3.09	2.89	0.0000
Pneumonia	3,218	3.08	2.85	0.0229
Heart Conduction	856	3.07	2.86	0.1991
Pancreatic Cancer	650	2.98	2.90	0.4105
Congestive Heart Failure	784	2.90	2.87	0.4746
Senility with Psychosis	121	2.85	2.86	0.5017
Heart Disease	11,524	2.82	2.89	0.9203
Alzheimer's Disease	232	2.81	3.00	0.5759

Table 13 (continued)

Cause of Death	N	Death Cases GIF	Control Cases GIF	P-Value*
Melanoma	236	2.76	2.85	0.5422
Esophageal Cancer	185	2.70	2.85	0.5492
Ulcer	399	2.63	2.94	0.6805
Bronchitis	254	2.54	2.76	0.6034
Intestinal Obstruction	174	2.50	3.06	0.6538
Motor Neuron Diseases	105	2.36	2.83	0.6001
Renal Failure	356	2.27	2.95	0.8319
Bladder Cancer	376	2.20	3.05	0.9042
Hodgkin's Disease	136	2.02	2.98	0.9042
Rectal Cancer	256	1.86	2.67	0.8559
Multiple Sclerosis	69	0.66	2.25	0.7761

*P-value for hypothesis that death cases GIF > control GIF

Table 14 GIF for Female Linked Death Certificates

Cause of Death	N	Death Cases GIF	Control Cases GIF	P-Value*
Hodgkin's Disease	**68**	**10.38**	**2.48**	**0.0004**
Multiple Sclerosis	**117**	**9.26**	**2.69**	**0.0000**
Kidney Cancer	**161**	**7.20**	**2.72**	**0.0002**
Aneurysm	**229**	**6.48**	**2.78**	**0.0001**
Mouth Cancers	69	5.66	2.39	0.1236
Asthma	159	5.30	2.66	0.0126
Parkinson's Disease	180	5.18	2.81	0.0253
Chronic Airway Obstruction	297	5.01	2.97	0.0073
Brain Cancer	290	5.00	2.70	0.0003
Obesity	148	5.00	2.83	0.0757
Nephritis	196	4.38	3.12	0.1114
Myeloma	239	4.24	2.90	0.0852
Motor Neuron Diseases	97	4.15	2.94	0.2735
Lung Cancer	**446**	**4.13**	**2.84**	**0.0084**
Diverticulosis	117	4.03	2.95	0.2737
Ulcer	281	3.92	2.96	0.1086
Emphysema	118	3.91	2.91	0.2835
Myeloid Leukemia	219	3.89	2.64	0.0905
Ovarian Cancer	738	3.87	2.84	0.0003
Cirrhosis	308	3.85	2.81	0.0583
Gallbladder Cancer	151	3.80	2.75	0.2117
Diabetes	3,454	3.71	2.89	0.0000
Cervical Cancer	166	3.53	2.89	0.3091
Pulmonary Embolism	339	3.52	2.80	0.1549
Lymphoma	446	3.46	2.85	0.0847
Senility Without Psych	400	3.38	2.72	0.1661
Renal Failure	319	3.35	3.07	0.3419
Heart Valve Disorders	475	3.30	2.86	0.1540
Congestive Heart Failure	1,269	3.25	3.04	0.1742
Colon Cancer	1,134	3.16	2.91	0.1379
Stroke	4,843	3.11	2.97	0.0305
Stomach Cancer	414	3.09	2.84	0.3306
Rectal Cancer	201	3.09	2.84	0.4057
Endocarditis	236	3.09	2.95	0.4376
Breast Cancer	2,203	3.03	2.81	0.0450
Circulatory Disorders	204	3.03	3.01	0.4918
Hypertension	3,078	3.01	2.88	0.0892
Cardiomyopathy	207	3.00	2.94	0.4777
Myocardial Infarction	4,831	2.98	2.94	0.2965
Pneumonia	3,139	2.87	2.91	0.6839
Uterine Cancer	416	2.86	2.97	0.5760
Heart Disease	8,956	2.84	2.92	0.9569
Motor Vehicle Accident	838	2.80	2.62	0.2297
Heart Conduction	926	2.75	3.00	0.8054
Alzheimer's Disease	332	2.71	2.89	0.6048
Intestinal Obstruction	309	2.70	2.95	0.6115

Table 14 (continued)

Cause of Death	N	Death Cases GIF	Control Cases GIF	P-Value*
Suicide	261	2.51	2.70	0.6141
Pancreatic Cancer	521	2.47	2.86	0.8077
Alcohol Related	75	2.39	3.27	0.6169
Bronchitis	120	2.37	2.68	0.5739
Lymphoid leukemia	110	2.29	2.91	0.6276
Congenital Anomalies	134	2.25	2.74	0.6309
Biliary Tract, Gallbladder	242	1.96	2.72	0.7849
Senility with Psychosis	225	1.95	3.10	0.8595
Melanoma	146	1.49	2.69	0.8365
Connective Tissue Cancer	81	1.02	3.24	0.7994
Influenza	143	0.86	2.44	0.8285
Liver Cancer	152	0.74	3.10	0.9212
Bladder Cancer	125	0.35	2.88	0.9098
Esophageal Cancer	43	0.32	2.99	0.7658

*P-value for hypothesis that death cases GIF > control GIF

Table 15 GIF Values for Male Death Certificates Compared to Female Death Certificates

Cause of Death	Male Cases GIF	Male Cases P-Value*	Female Cases GIF	Female Cases P-value*
Males Higher				
Influenza	7.57	0.0131	0.86	0.8285
Congenital Anomalies	7.47	0.0000	2.25	0.6309
Gallbladder Cancer	6.34	0.0233	3.80	0.2117
Connective Tissue Cancer	5.97	0.1275	1.02	0.7994
Alcohol Related	**5.44**	**0.0000**	**2.39**	**0.6169**
Myeloid Leukemia	4.98	0.0011	3.89	0.0905
Circulatory Disorders	4.90	0.0303	3.03	0.4918
Liver Cancer	4.83	0.0688	0.74	0.9212
Biliary Tract and Gallbladder Disorders	4.17	0.1203	1.96	0.7849
Suicide	**4.02**	**0.0000**	**2.51**	**0.6141**
Endocarditis	3.90	0.1120	3.09	0.4376
Pulmonary Embolism	3.85	0.0167	3.52	0.1549
Stomach Cancer	3.44	0.0366	3.09	0.3306
Stroke	3.41	0.0000	3.11	0.0305
Hypertension	3.25	0.0111	3.01	0.0892
Myocardial Infarction	3.09	0.0000	2.98	0.2965
Pneumonia	3.08	0.0229	2.87	0.6839
Females Higher				
Hodgkin's Disease	2.06	0.0844	10.38	0.0004
Multiple Sclerosis	0.66	0.7761	9.26	0.0000
Kidney Cancer	5.90	0.0000	7.20	0.0002
Aneurysm	3.51	0.0604	6.48	0.0001
Asthma	3.87	0.2243	5.30	0.0126
Parkinson's Disease	3.77	0.0796	5.18	0.0253
Chronic Airway Obstruct.	3.41	0.0508	5.01	0.0073
Brain Cancer	3.17	0.1983	5.00	0.0003
Obesity	4.16	0.3090	5.00	0.0757
Lung Cancer	**3.28**	**0.0040**	**4.13**	**0.0084**
Ulcer	2.63	0.6805	3.92	0.1986
Cirrhosis	3.47	0.1589	3.85	0.0583
Congestive Heart Failure	2.90	0.4746	3.25	0.1742
Rectal Cancer	1.86	0.8559	3.09	0.4057

*P-value for hypothesis that death cases GIF > control GIF

Table 16 Two-Way GIF for Linked Death Certificates

Cause of Death (one-way GIF)	N	Combined GIF	Combined Controls GIF	P-value
Breast Cancer(3.03)				
Ovarian Cancer(3.87)	2,941	2.98	2.81	0.0457
Prostate Cancer(3.88)	4,684	3.23	2.85	0.0000
Colon Cancer(3.14)				
Prostate Cancer(3.88)	4,727	3.25	2.88	0.0000
Prostate Cancer(3.88)				
Stomach Cancer(3.56)	3,470	3.41	2.88	0.0000
Bladder Cancer(2.62)	2,982	3.54	2.88	0.0000
Ovarian Cancer(3.87)				
Uterine Cancer(2.86)	1,154	3.71	2.85	0.0001
Lymphoid Leukemia(3.95)				
Myeloid Leukemia(3.86)	753	3.70	2.76	0.0010
Connect. Tissue Ca.(2.49)	**455**	**4.30**	**2.83**	**0.0005**
Myeloid Leukemia(3.86)				
Connect. Tissue Ca.(2.49)	**624**	**3.92**	**2.81**	**0.0008**
Hodgkin's Disease(4.95)				
Lymphoma(3.67)	1,173	3.25	2.87	0.0261
Diabetes(3.64)				
Renal Failure(2.80)	6,663	3.51	2.88	0.0000
Congestive Heart(3.28)	6,504	3.53	2.86	0.0000
Parkinson's Disease(3.91)				
Motor Neuron Diseases(4.16)	707	3.52	2.82	0.0074
Multiple Sclerosis(6.86)	691	3.54	2.88	0.0526
Heart Disease(2.81)				
Myocardial Infarction(3.07)	**30,002**	**3.14**	**3.03**	**0.0000**
Pulmonary Embolism(3.57)				
Chronic Airway Obst.(3.95)	1,994	3.50	2.88	0.0000
Asthma(3.97)				
Chronic Airway Obst.(3.95)	1,579	3.62	2.90	0.0000
Emphysema(3.72)	1,481	3.65	2.94	0.0002
Chronic Airway Obst.(3.95)				
Emphysema(3.72)	2,458	3.64	2.95	0.0000

combined GIF is higher than the single GIF for either cause of death. This occurred with connective tissue cancer when it was combined with lymphoid leukemia and myeloid leukemia. A higher combined GIF was also seen for the combination of heart disease and myocardial infarction. Many of the combined GIF scores have low p-values, but usually both of the diseases had a low p-value by themselves. Significant values are seen for almost every combination with diabetes. This can be attributed mostly to the diabetes set since it is so much larger than most of the other sets. The diabetes set by itself has a large GIF and a large sample size. The GIF is calculated by adding the kinship coefficient for each pair and dividing by the total number of pairs. When diabetes is combined with other causes of death, the kinship coefficient for the diabetes cases is still large enough to produce a large GIF value, but it is divided by a larger number of pairs. The contribution to the GIF from the two causes of death is not enough to offset the increased number of cases, so the combined GIF is lower.

An important quality of the index of familiality is that it is calculated using both close and distant relatives. In order to determine whether a high GIF value was reflective of familial influence or genetic predisposition, the contribution to the GIF by path length was examined. Close relatives such as siblings would have a path of two, whereas distant relatives would have a path of 9 or 10 individuals between them. The contribution to the GIF by path length was plotted for 10 of the causes of death. Ten plots were done to get a sample of different types of causes of death. Some of the causes of death plotted such as motor vehicle accidents were obviously not genetic, but the graphs illustrate the familiality. Other causes of death such as diabetes were plotted because there is genetic evidence for diabetes and the GIF score should be a result of relationships among both close relatives and more distant relatives. The contribution to the GIF for the death cases is drawn, along with the contribution to the GIF for the median control value and the 5th and 95th percentile GIFs for the control group. The plots are the most interesting where there are gaps between the line for the death cases and the lines for the control cases, especially where the path lengths are larger. In these cases, a genetic cause is likely since distant relatives would not share the same environment.

Figure 18 covers lung cancer and emphysema. Both of these diseases are related to smoking and could be a result of families sharing the same environment. For lung cancer, the contribution to the GIF is the same for the death cases and the controls after only three lengths, suggesting that lung cancer is mostly familial and not genetic. The death certificates line for emphysema is higher than the controls median line for six path lengths, suggesting a genetic predisposition.

Figure 19 shows plots for influenza and pneumonia. The GIF for influenza was high, but there was a small sample size. The small sample size produces a plot with large variation of the contribution values that is difficult to interpret. For pneumonia, there is no difference between the control lines and the death cases line. There was a large sample set for pneumonia, so there is not a lot of

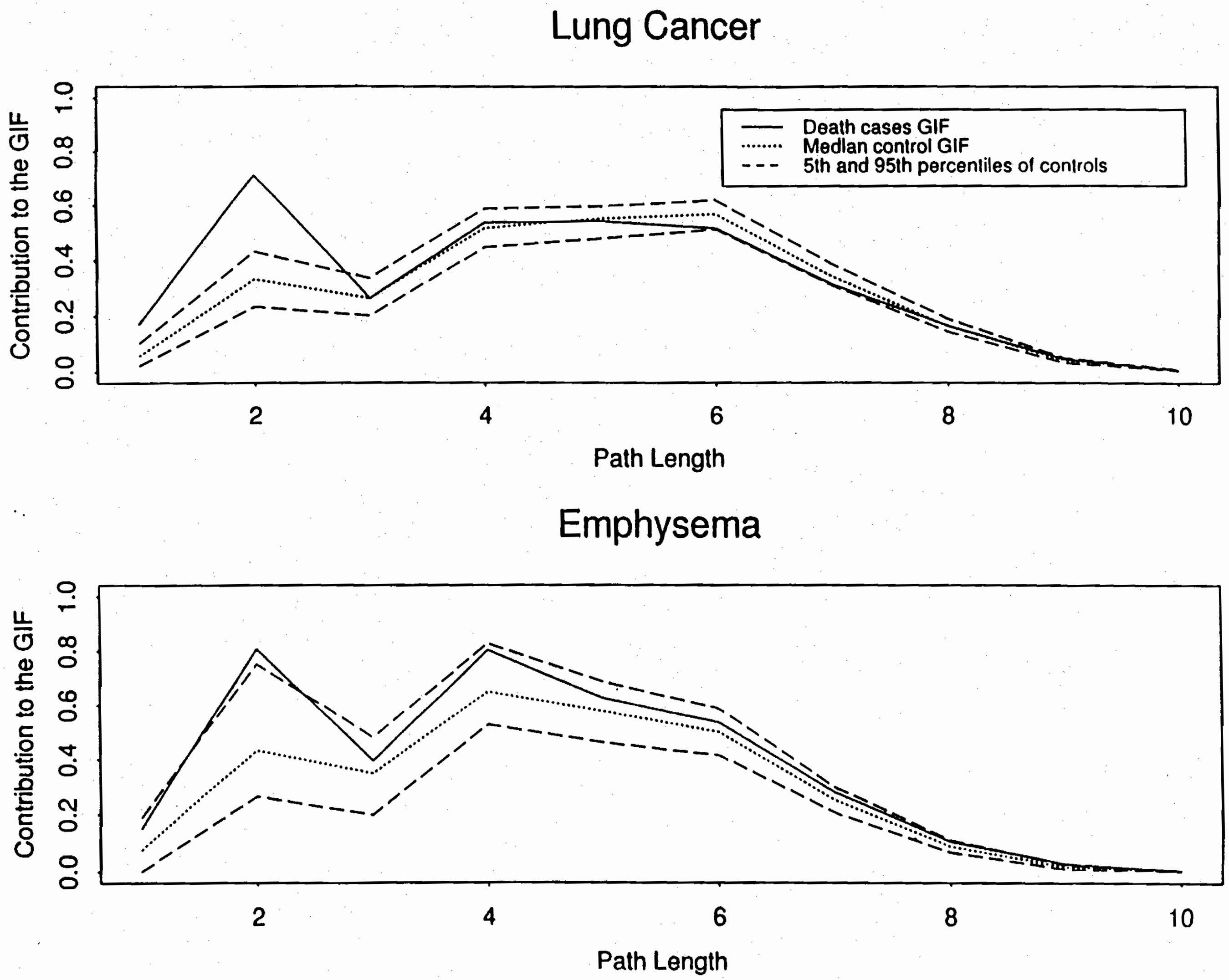


Figure 18 Contribution to the GIF by Path Length - Lung Cancer and Emphysema

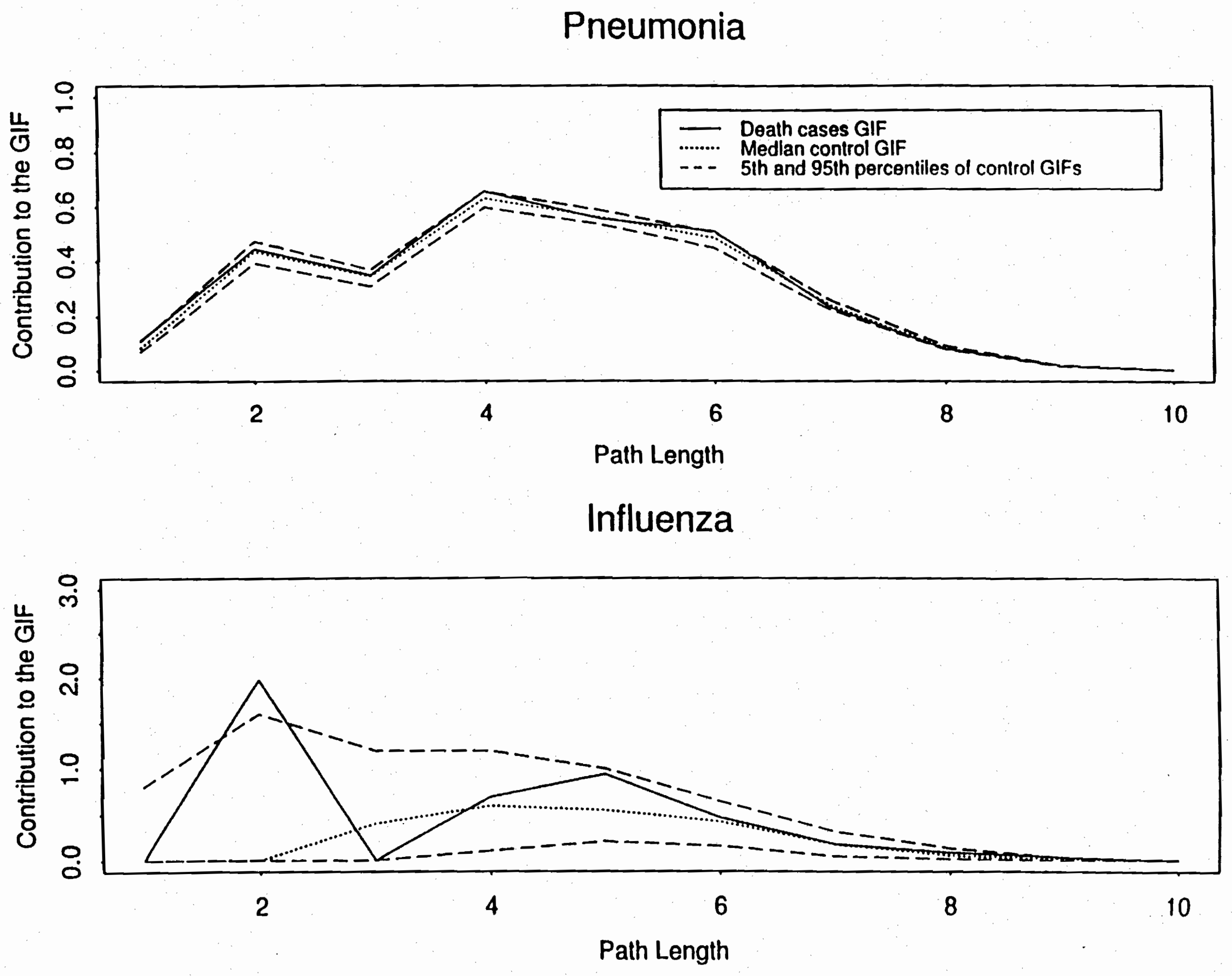


Figure 19 Contribution to the GIF by Path Length - Pneumonia and Influenza

variation in the control lines.

Figure 20 shows suicide and motor vehicle accidents. Suicide has significant contributions to the death cases GIF in the first two path lengths, and motor vehicle accidents has a similar death case line with a smaller gap between the controls and the death cases. After two path lengths, the death case GIF is the same as the controls. Both causes of death seem to be familial and not genetic. This is expected from motor vehicle accidents, since it is common for more than one family member to die in the same accident.

Kidney cancer and multiple sclerosis had some of the highest GIF values. In Figure 21, both causes of death seem to have some genetic influence with higher contributions to the death case GIF for the fourth and fifth path lengths. These plots are scaled differently from the others. The contribution to the GIF goes from 0 to 3, rather than 0 to 1 like the other plots. This was done to show the large contribution to the GIF for multiple sclerosis in the second, third, and fourth path lengths.

The plots for diabetes and chronic airway obstruction are shown in Figure 22. A genetic influence is shown for both causes of death, but it is only evident for four path lengths.

4.4 First-Degree Relative Risk Methods

As with the GIF study, a study of the first-degree relative risk with the linked cancer records has been done previously (Goldgar 1994). Using the

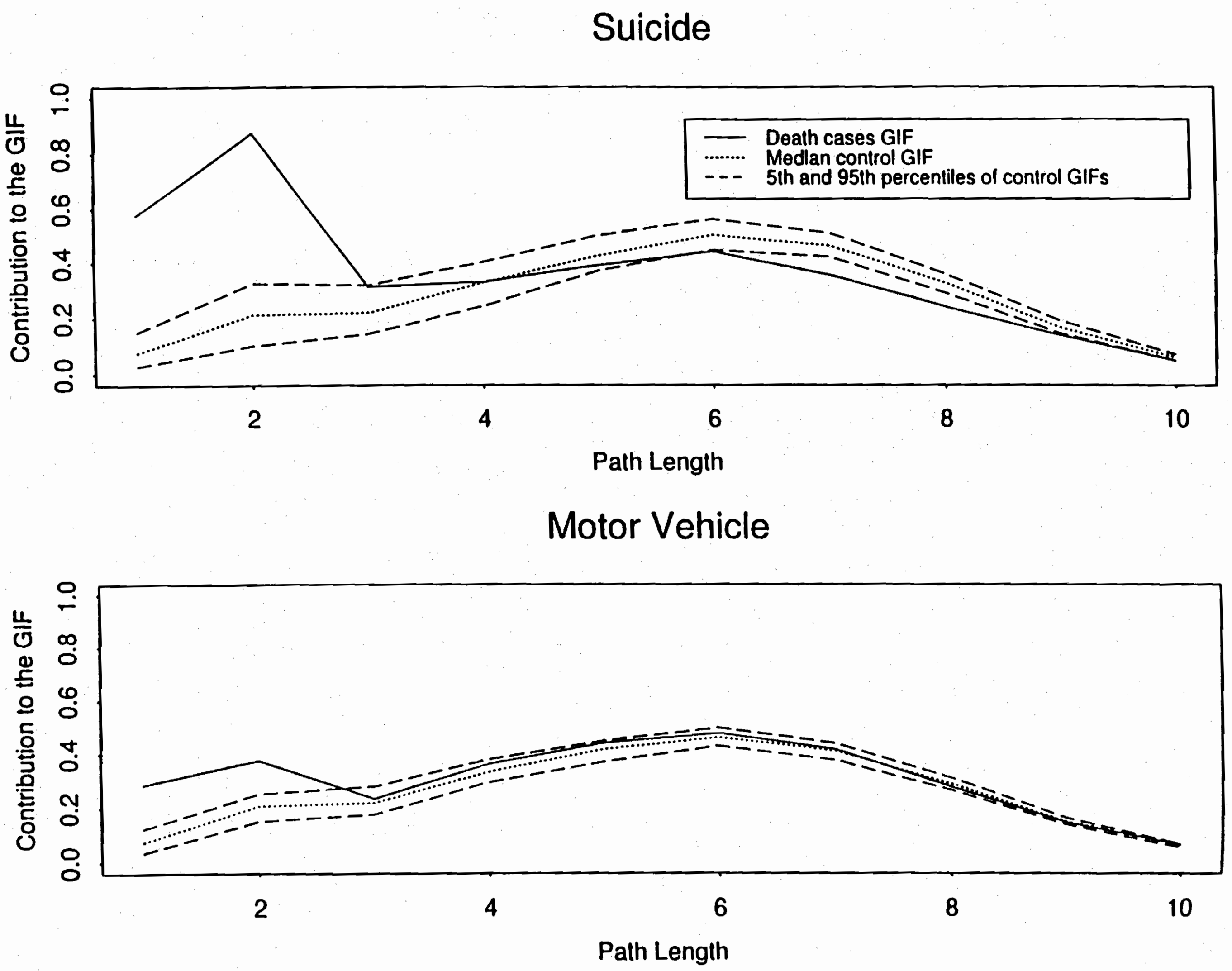
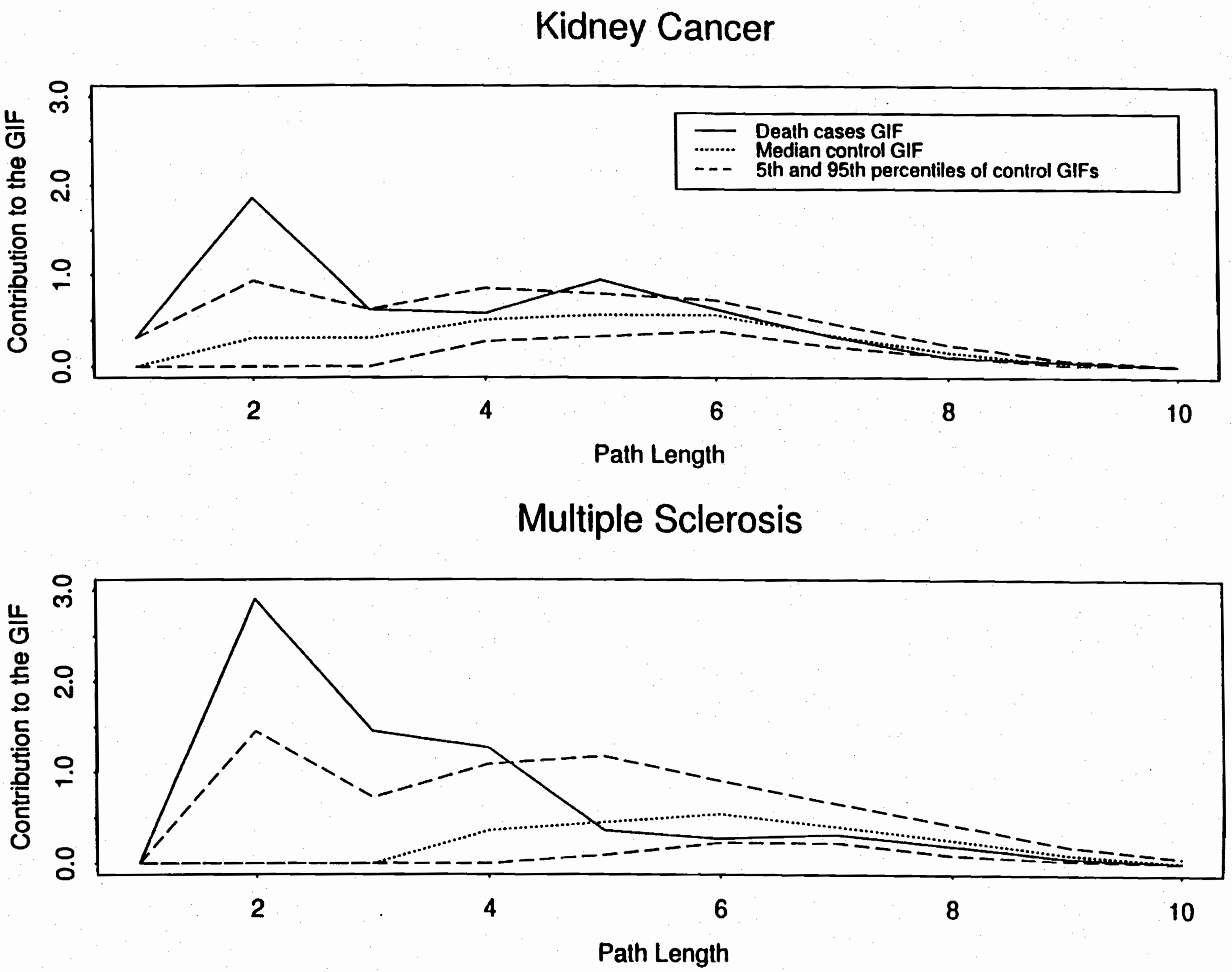


Figure 20 Contribution to the GIF by Path Length - Suicide and Motor Vehicle


Kidney Cancer

Multiple Sclerosis

Figure 21 Contribution to the GIF by Path Length - Kidney Cancer and Multiple Sclerosis

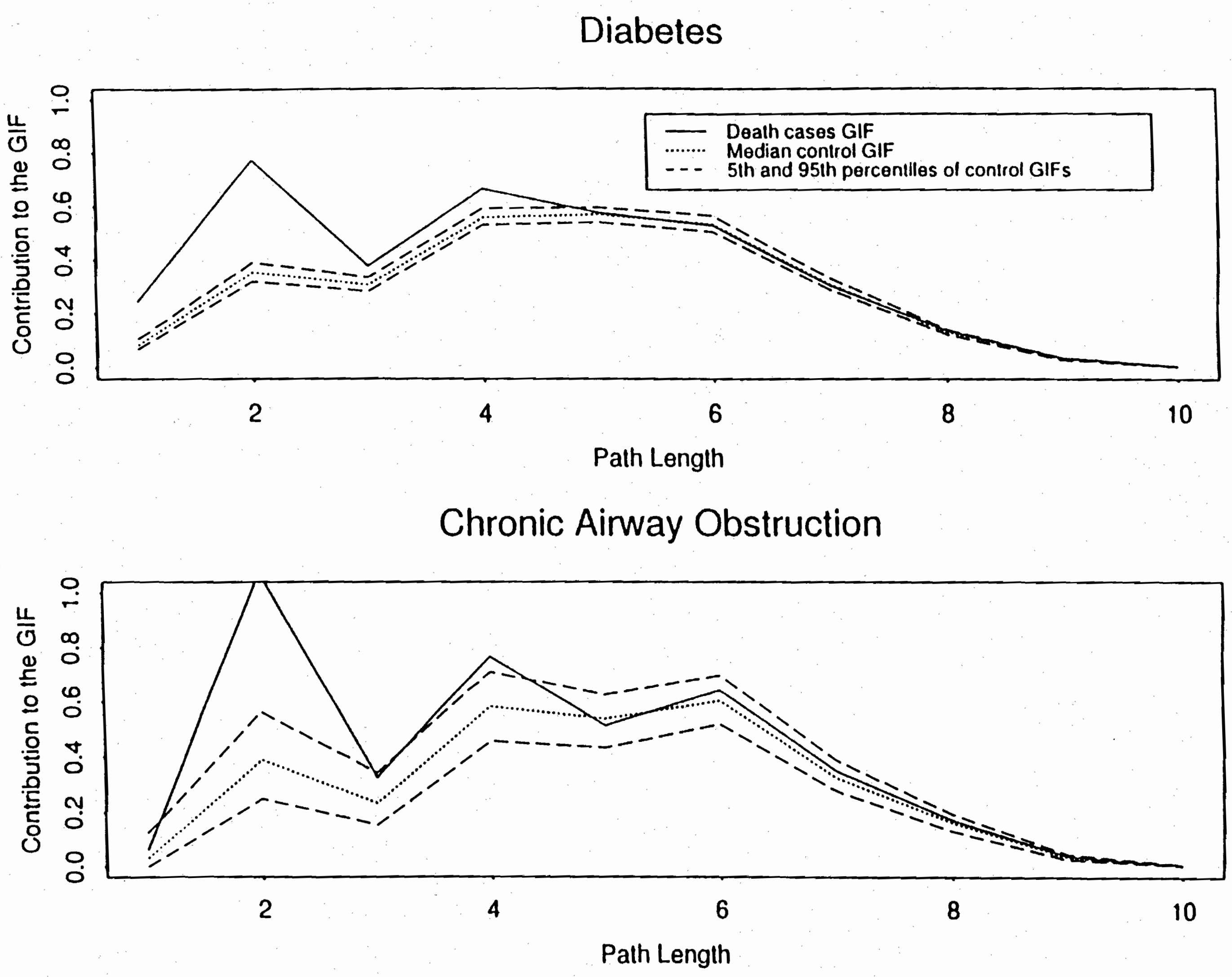


Figure 21 Contribution to the GIF by Path Length - Diabetes and Chronic Airway Obstruction

method from this study, a similar analysis of the linked death certificates was done. The cancer results were compared with the results from the linked cancer records study.

The first step in the calculation of the first-degree relative risk was to use the genealogic information available in the UPDB to gather all of the first-degree relatives of each individual in each cause of death group. A first-degree relative is a sibling, parent, or offspring. If any of the first-degree relatives were part of the linked death certificates, the cause of death and age of death were added to the data file. The death certificate cases in each group were marked as probands.

Relatives of the probands were divided into 64 cohort groups based on year of birth, sex, and whether they were born in Utah or outside of Utah. The total number of individuals in each cohort was calculated. The 64 cohort groups were further divided by cause of death and the decade of age at death. The categories for age of death used were less than 40, 40-49, 50-59, 60-69, 70-79, greater than 80, and unknown. This resulted in the total number of individuals in each of the 64 cohort groups with 61 X 7 entries for each cohort.

The control group for this study was selected from the first-degree relatives of all the individuals in the UPDB who were known to have died in Utah. Only those death certificates where the individual died in Utah were used in the record linking. The assumption was made that on average, the mortality and migration experience for a given cohort of the relatives of the death cases was the same as the equivalent control cohort of relatives of individuals who had died in Utah. This insured that there would not be a misrepresentation of individuals in the control group who had migrated out of Utah and thus could not be in the death certificate links. The first-degree relatives of all of those individuals who had died in Utah were collected. Cause of death and age of death were added to the first degree relative record when known. The control file was also put into the same cohort format as the case files.

The control individuals were used to obtain internal cohort specific rates that are used to calculate the expected value for each cause of death. This is calculated by

$$E = \sum_{i=1}^{64} R_i{*}C_i/N_i$$

where R_i is the number of relatives of the probands, N_i is the number of controls and C_i is the number of cases of each cause of death found among these control subjects in the *i*th control group. The estimated relative risk can be calculated by dividing the number of observed cases for each cause of death by the expected value for each cause of death. The estimated relative risk was calculated using these observed and expected numbers for each of the seven age categories and then combined to produce the overall relative risk.

Approximate 95% confidence limits and hypothesis tests of the null hypothesis of relative risk = 1 can be calculated, assuming the numbers of cases of cause of death found among the relatives follow a Poisson distribution with mean O (observed). Because of the large number of calculations needed to calculate exact Poisson probabilities in this study, a normal approximation was used. The limit of at least 100 cases in each cause of death set makes this approximation feasible. The transformation

$$Z=2\left(\sqrt{O}-\sqrt{E}\right)$$

was used to approximate a normal distribution (Miettinen 1985) . Under the null hypothesis that familial relative risk = 1, Z has mean 0 and variance 1. These Z values can be used to compute P values for testing the null hypothesis and 95% confidence limits for familial relative risk. The values

$$O_L = \left(\sqrt{O}-0.98\right)^2 \qquad O_U = \left(\sqrt{O}+0.98\right)^2$$

are divided by the expected number of cases (E) to estimate the lower and upper 95% confidence limits for familial relative risk. Since the relative risk was most interesting when it was greater than one, a one-sided p-value using the approximated Z value was calculated to test the hypothesis that the relative risk was greater than one. This is consistent with the one-sided p-value that was calculated for the GIF.

4.5 First-Degree Relative Risk Results

The 61 cause of death groups were used in the relative risk study. The familial relative risk was calculated for each group. Familial relative risk values were also calculated for the youngest third of each group and the males and females in each group. A comparison with the previous results from the cancer registry study was done.

If a cause of death had less than four first-degree relatives with that cause of death, the relative risk was either less than one or the confidence interval was large. The confidence intervals ranged far below one and often went to zero. This was common in causes of death where there was a small sample size such as multiple sclerosis or Hodgkin's disease. Since the reliability of the relative risk values was low for these sets, they were not included in the results table. There were also causes of death where there were no first-degree relatives with the same cause of death, so those sets are not in the table.

Table 17 shows the results for all of the causes of death that had at least four first-degree relatives with the same cause of death. They are ordered by relative risk score. The highest values come from alcohol related, kidney cancer, and mouth cancer; however all of these causes of death had large confidence intervals. Because the confidence interval implies a 95% confidence that the correct value will be within the interval, a smaller confidence interval increases the reliability of the estimated relative risk. Causes of death with a high relative risk and a small confidence interval are suicide, aneurysm, chronic airway obstruction, and prostate cancer. Relative risk scores of one or lower come from pancreatic cancer, heart conduction disorders, and pneumonia. The causes of death with the largest sample sizes had the smallest confidence intervals and

Table 17 First-Degree Relative Risk for All Death Cases

Cause of Death	N	Observed	Expected	Relative Risk	95% CI	P-value*
Alcohol Related	**390**	**13**	**2.12**	**6.13**	**2.32-11.74**	**.0012**
Kidney Cancer	**391**	**13**	**2.14**	**6.08**	**2.30-11.66**	**.0012**
Mouth Cancers	**264**	**6**	**0.99**	**6.07**	**1.14-14.88**	**.0197**
Gallbladder Cancer	228	4	0.68	5.89	0.56-16.89	.0485
Congenital Anomalies	264	8	1.47	5.42	1.41-12.04	.0113
Myeloma	531	17	3.56	4.78	2.11-8.53	.0008
Suicide	**1,411**	**108**	**23.29**	**4.64**	**3.48-5.96**	**.0000**
Myeloid Leukemia	445	9	2.24	4.01	1.16-8.58	.0170
Aneurysm	**861**	**34**	**8.79**	**3.87**	**2.25-5.93**	**.0000**
Circulatory Disorder	346	5	1.41	3.53	0.51-9.27	.0694
Parkinson's Disease	473	9	2.61	3.45	1.00-7.36	.0250
Nephritis	457	9	2.79	3.23	0.93-6.89	.0301
Chronic Airway	**1,214**	**58**	**19.48**	**2.98**	**1.99-4.16**	**.0000**
Prostate Cancer	**2,259**	**167**	**59.20**	**2.82**	**2.25-3.46**	**.0000**
Alzheimer's Disease	471	6	2.14	2.80	0.53-6.86	.0823
Cardiomyopathy	393	5	1.83	2.74	0.40-7.19	.1056
Ovarian Cancer	715	17	6.62	2.57	1.13-4.58	.0143
Diabetes	**5,524**	**852**	**357.41**	**2.38**	**2.16-2.62**	**.0000**
Pulmonary Embolism	666	13	5.57	2.33	0.88-4.47	.0392
Lung Cancer	**2,046**	**134**	**61.23**	**2.19**	**1.70-2.74**	**.0000**
Motor Vehicle	2,452	143	68.37	2.09	1.63-2.60	.0000
Lymphoma	928	22	10.25	2.15	1.07-3.60	.0174
Cirrhosis	692	14	6.97	2.01	0.80-3.77	.0594
Brain Cancer	584	9	4.50	2.00	0.58-4.27	.0384
Heart Valve Disorders	840	21	10.67	1.97	0.96-3.34	.0314
Breast Cancer	**2,084**	**102**	**54.35**	**1.88**	**1.40-2.43**	**.0000**
Biliary, Gallbladder	413	4	2.16	1.85	0.48-4.12	.1446
Stomach Cancer	937	20	11.58	1.73	0.82-2.96	.0655
Emphysema	1,138	37	21.82	1.70	1.01-2.56	.0228
Ulcer	623	8	5.11	1.56	0.41-3.47	.2119
Colon Cancer	2,085	86	56.03	1.53	1.11-2.03	.0057
Myocardial Infarct.	12,570	2,984	2043.41	1.46	1.39-1.54	.0000
Hypertension	4,311	280	192.77	1.45	1.22-1.70	.0000
Congestive Heart	1,502	23	16.98	1.35	0.68-2.25	.1685
Stroke	6,785	567	436.87	1.30	1.15-1.45	.0000
Heart Disease	17,714	4,949	4058.30	1.22	1.17-1.27	.0000
Pneumonia	4,957	206	205.19	1.00	0.82-1.21	.4840
Heart Conduction	1,478	20	21.79	0.92	0.44-1.58	.6103
Pancreatic Cancer	1,104	14	15.53	0.90	0.36-1.69	.6103

*one-sided p-value for hypothesis that relative risk > 1

therefore the most reliable relative risks. These included breast cancer, lung cancer, and diabetes.

For those diseases with a large enough sample size that would produce meaningful results, the relative risk was calculated for the youngest one-third of each set. The results are in Table 18. In most cases, the youngest set had a higher relative risk than the whole set. Large differences were seen in myeloma, aneurysm, nephritis, diabetes, and prostate cancer. Those causes of death were the youngest relative risk was higher are listed in Table 19. There were some causes of death where the youngest relative risk was lower such as myeloid leukemia, ovarian cancer, lung cancer, and heart valve disorders.

Tables 20 and 21 show the relative risk for the causes of death for the males and females in each set where there were enough cases. Significant relative risk value for males were seen in suicide, prostate cancer, diabetes, motor vehicle accidents, and lung cancer. Increased risk for females is shown for chronic airway obstruction, lung cancer, diabetes, and breast cancer. The causes of death where there are substantial differences between males and females are listed in Table 22. The risks for colon cancer and pulmonary embolisms were almost twice as high for males as compared to females. The largest increases in risk for females were with aneurysm, nephritis, chronic airway obstruction, and ulcers.

The last analysis of the relative risk was to look at the relative risk for other causes of death than the proband's. For example, the risk of prostate cancer where the probands are breast cancer deaths was examined. The results of this

Table 18 First-Degree Relative Risk for Youngest Third of Death Cases

Cause of Death				Relative		
(oldest age)	N	Observed	Expected	Risk	95% CI	P-value*
Myeloma(70)	**286**	**12**	**1.71**	**7.01**	**2.52-13.74**	**.0011**
Aneurysm(70)	**321**	**20**	**2.98**	**6.70**	**3.19-11.50**	**.0000**
Kidney Cancer(65)	143	4	0.65	6.17	0.58-17.68	.0455
Alcohol Related(60)	185	6	0.97	6.16	1.16-15.10	.0192
Nephritis(65)	**185**	**5**	**0.85**	**5.87**	**0.85-15.41**	**.0314**
Suicide(40)	458	108	22.48	4.80	3.61-6.17	.0000
Prostate Cancer(73)	**734**	**71**	**16.72**	**4.25**	**2.96-5.76**	**.0000**
Parkinson's(75)	183	4	1.00	4.00	0.38-11.43	.0793
Myeloid Leukemia(65)	188	3	0.78	3.86	0.15-12.50	.1151
Chronic Airway(70)	348	20	5.24	3.82	1.82-6.55	.0010
Diabetes(70)	**1,782**	**333**	**95.39**	**3.49**	**2.98-4.04**	**.0000**
Pulmonary Emb.(70)	290	7	2.05	3.41	0.77-7.93	.0427
Lymphoma(67)	319	10	3.08	3.25	1.02-6.71	.0234
Alzheimer's(80)	258	4	1.23	3.24	0.31-9.30	.1038
Emphysema(65)	324	17	5.53	3.07	1.35-5.48	.0062
Motor Vehicle(40)	980	73	25.58	2.85	2.00-3.85	.0000
Ulcer(70)	257	5	1.85	2.71	0.39-7.10	.1075
Ovarian Cancer(60)	214	4	1.69	2.36	0.22-6.77	.1611
Brain Cancer(60)	311	5	2.15	2.33	0.34-6.11	.1379
Hypertension(73)	1,473	123	56.40	2.18	1.67-2.76	.0000
Lung Cancer(64)	658	38	17.81	2.13	1.28-3.20	.0030
Breast Cancer(60)	815	39	18.45	2.11	1.28-3.16	.0029
Cirrhosis(60)	275	5	2.49	2.01	0.29-5.28	.1762
Stomach Cancer(67)	340	7	3.50	2.00	0.45-4.64	.1379
Congestive Heart(79)	582	12	6.30	1.90	0.69-3.73	.0885
Colon Cancer(66)	650	27	14.78	1.83	0.98-2.93	.0281
Myocardial Infct(69)	4,384	1,154	640.02	1.80	1.66-1.95	.0000
Heart Valve(60)	359	7	4.03	1.74	0.39-4.03	.1841
Heart Disease(71)	6,005	1,810	1,196.06	1.51	1.42-1.61	.0000
Stroke(75)	2,047	166	120.09	1.38	1.10-1.70	.0033
Pancreatic Cancer(65)	346	5	3.97	1.26	0.18-3.30	.3669
Pneumonia(75)	1,549	60	57.17	1.05	0.71-1.46	.3974

*one-sided p-value for hypothesis that relative risk > 1

Table 19 First-Degree Relative Risk for All Death Cases Compared to Youngest Third

	All Cases		Youngest Third	
Cause of Death	Relative Risk	95% CI	Relative Risk	95% CI
Myeloma	4.78	2.11-8.53	7.01	2.52-13.74
Aneurysm	3.87	2.25-5.93	6.70	3.19-11.50
Nephritis	3.23	0.93-6.89	5.87	0.85-15.41
Chronic Airway Obst.	2.98	1.99-4.16	3.82	1.82-6.55
Prostate Cancer	2.82	2.25-3.46	4.25	2.96-5.76
Diabetes	2.38	2.16-2.62	3.49	2.98-4.04
Pulmonary Embolism	2.33	0.88-4.47	3.41	0.77-7.93
Lymphoma	2.15	1.07-3.60	3.25	1.02-6.71
Motor Vehicle	2.09	1.63-2.60	2.85	2.00-3.85
Breast Cancer	1.88	1.40-2.43	2.11	1.28-3.16
Stomach Cancer	1.73	0.82-2.96	2.00	0.45-4.64
Emphysema	1.70	1.01-2.56	3.07	1.35-5.48
Ulcer	1.56	0.41-3.47	2.71	0.39-7.10
Colon Cancer	1.53	1.11-2.03	1.83	0.98-2.93
Myocardial Infarct.	1.46	1.39-1.54	1.80	1.66-1.95
Hypertension	1.45	1.22-1.70	2.18	1.67-2.76
Congestive Heart	1.35	0.68-2.25	1.90	0.69-3.73
Heart Disease	1.22	1.17-1.27	1.51	1.42-1.61

Table 20 First-Degree Relative Risk for Male Death Cases

Cause of Death	N	Observed	Expected	Relative Risk	95% CI	P-value*
Alcohol Related	315	13	1.74	7.46	2.83-14.30	.0006
Kidney Cancer	240	8	1.27	6.28	1.07-5.82	.0082
Suicide	**1,151**	**93**	**19.53**	**4.76**	**3.49-6.23**	**.0000**
Myeloma	306	10	2.14	4.66	1.47-9.65	.0082
Myeloid Leukemia	237	5	1.14	4.40	0.64-11.54	.0495
Parkinson's Disease	308	6	1.73	3.48	0.66-8.53	.0537
Aneurysm	663	22	6.72	3.27	1.62-5.49	.0015
Pulmonary Embolism	361	9	2.96	3.04	0.88-6.49	.0351
Prostate Cancer	**2,259**	**167**	**59.20**	**2.82**	**2.25-3.46**	**.0000**
Diabetes	**2,404**	**372**	**147.34**	**2.52**	**2.17-2.90**	**.0000**
Heart Valve	383	12	4.81	2.50	0.90-4.89	.0359
Chronic Airway	926	35	14.51	2.41	1.41-3.67	.0014
Nephritis	276	4	1.66	2.41	0.23-6.92	.1562
Alzheimer's Disease	199	2	0.85	2.34	0.00-9.19	.2451
Brain Cancer	346	6	2.61	2.30	0.43-5.63	.1190
Motor Vehicle	**1,642**	**94**	**45.42**	**2.07**	**1.52-2.70**	**.0000**
Lung Cancer	**1,625**	**98**	**47.78**	**2.05**	**1.52-2.67**	**.0000**
Colon Cancer	1,047	52	27.31	1.90	1.24-2.71	.0025
Lymphoma	505	10	5.37	1.86	0.59-3.85	.1151
Emphysema	1,024	34	19.78	1.72	1.00-2.63	.0250
Stomach Cancer	553	11	6.60	1.67	0.56-3.35	.1446
Cirrhosis	390	6	3.87	1.55	0.29-3.80	.2483
Congestive Heart	625	10	6.83	1.46	0.46-3.03	.2206
Myocardial Inf.	8,294	1,896	1,312.62	1.44	1.35-1.54	.0000
Stroke	2,939	260	186.40	1.39	1.17-1.64	.0002
Hypertension	1,721	101	72.81	1.39	1.03-1.80	.0158
Heart Disease	10,502	2,860	2,308.91	1.24	1.18-1.30	.0000
Pancreatic Cancer	620	9	8.50	1.06	0.31-2.26	.4522
Pneumonia	2,678	107	108.64	0.98	0.74-1.27	.5438
Ulcer	371	2	3.00	0.67	0.00-2.61	.6736

*one-sided p-value for hypothesis that relative risk > 1

Table 21 First-Degree Relative Risk for Female Death Cases

				Relative		
Cause of Death	N	Observed	Expected	Risk	95% CI	P-value*
Aneurysm	198	12	2.07	5.81	2.09-11.39	.0021
Kidney Cancer	151	5	0.86	5.80	0.84-15.21	.0322
Myeloma	225	7	1.41	4.96	1.12-11.51	.0197
Chronic Airway	**288**	**23**	**4.97**	**4.63**	**2.34-7.69**	**.0001**
Nephritis	181	5	1.13	4.41	0.64-11.57	.0485
Suicide	260	15	3.75	4.00	1.65-7.37	.0031
Myeloid Leukemia	208	4	1.11	3.62	0.34-10.37	.0901
Parkinson's Disease	165	3	0.89	3.38	0.14-10.96	.1314
Alzheimer's Disease	272	4	1.29	3.10	0.29-8.88	.1112
Ulcer	252	6	2.11	2.84	0.54-6.96	.0792
Lung Cancer	**421**	**36**	**13.45**	**2.68**	**1.58-4.06**	**.0005**
Brain Cancer	284	6	2.29	2.61	0.49-6.41	.0934
Cirrhosis	302	8	3.10	2.58	0.67-5.73	.0655
Ovarian Cancer	715	17	6.62	2.57	1.13-4.58	.0143
Lymphoma	423	12	4.88	2.46	0.89-4.82	.0375
Diabetes	**3,120**	**480**	**210.07**	**2.28**	**2.01-2.58**	**.0000**
Motor Vehicle	810	49	22.95	2.14	1.37-3.06	.0009
Breast Cancer	**2,084**	**102**	**54.35**	**1.88**	**1.40-2.43**	**.0000**
Stomach Cancer	384	9	4.98	1.81	0.52-3.86	.1379
Pulmonary Embolism	305	4	2.61	1.53	0.14-4.39	.2946
Heart Valve	457	9	5.87	1.53	0.44-3.28	.2061
Myocardial Inf.	4,276	1,088	730.78	1.49	1.37-1.62	.0000
Hypertension	2,590	179	119.95	1.49	1.20-1.82	.0003
Emphysema	114	3	2.03	1.47	0.06-4.78	.3336
Congestive Heart	877	13	10.14	1.28	0.49-2.46	.2743
Stroke	3,846	307	250.48	1.23	1.04-1.43	.0082
Colon Cancer	1,038	34	28.73	1.18	0.69-1.81	.2514
Heart Disease	7,212	2,089	1,749.39	1.19	1.12-1.27	.0000
Pneumonia	2,279	99	96.56	1.03	0.76-1.33	.4325
Pancreatic Cancer	347	2	4.86	0.41	0.00-1.61	.8686

*one-sided p-value for hypothesis that relative risk >

Table 22 First-Degree Relative Risk for Male Death Cases Compared to Female Death Cases

	Male Cases		Female Cases	
Cause of Death	Relative Risk	95% CI	Relative Risk	95% CI
Males Higher				
Suicide	4.76	3.49-6.23	4.00	1.65-7.37
Myeloid Leukemia	4.40	0.64-11.54	3.62	0.34-10.37
Heart Valve	2.50	0.90-4.89	1.53	0.44-3.28
Pulmonary Embolism	**3.04**	**0.88-6.49**	**1.53**	**0.14-4.39**
Heart Valve	2.50	0.90-4.89	1.53	0.44-3.28
Colon Cancer	**1.90**	**1.24-2.71**	**1.18**	**0.69-1.81**
Females Higher				
Aneurysm	**3.27**	**1.62-5.49**	**5.81**	**2.09-11.39**
Nephritis	**2.41**	**0.23-6.92**	**4.41**	**0.64-11.57**
Chronic Airway	**2.41**	**1.41-3.67**	**4.63**	**2.34-7.69**
Alzheimer's disease	2.34	0.00-9.19	3.10	0.29-8.88
Lung Cancer	2.05	1.52-2.67	2.68	1.58-4.06
Cirrhosis	1.55	0.29-3.80	2.58	0.67-5.73
Ulcer	**0.67**	**0.00-2.61**	**2.84**	**0.54-6.96**
Lymphoma	1.86	0.59-3.85	2.46	0.89-4.82

analysis are shown in Table 23. Only those causes of death that had a familial relative risk greater than 1.4 or a p-value less than .01 were included in the table. For some of the causes of death, there were no significant relative risk scores with other causes of death.

Associations were seen between alcohol-related deaths and mouth cancers along with suicide and alcohol-related deaths. A number of associations were seen among the different types of cancer deaths such as breast-ovarian and colon-prostate. Almost every cancer was associated with at least one other type of cancer. Significant associations with emphysema were observed for bronchitis, chronic airway obstruction, and lung cancer. Associations were also seen for the risk factors of heart disease and obesity for myocardial infarctions.

4.6 Comparison of GIF Results to Relative Risk Results

Both the genealogical index and the first-degree relative risk help to find excess familiality. The genealogical index does have the advantage of examining more of the extended family relationships, since it looks beyond first-degree relatives. An excess of a disease in distant relatives is a good indication of genetic predisposition, because environmental factors would have a smaller effect. A cause of death that ranked high for both analysis most likely has a genetic predisposition, whereas a cause of death that ranked high only in the first-degree relative risk could be strictly familial and not have a genetic predisposition.

In order to compare the results from the two methods, Table 24 was made

Table 23 Relative Risk for Related Causes of Death

Cause of Death	Observed	Expected	Relative Risk	95% CI	P-value
Alcohol Related					
Mouth Cancer	**5**	**1.25**	**4.01**	**1.26-8.29**	**.0125**
Aneurysm	9	3.25	2.77	1.26-4.88	.0082
Chronic Airway Obst.	12	5.45	2.20	1.13-3.62	.0119
Ulcer	5	2.31	2.17	0.68-4.48	.0764
Stomach Cancer	7	3.78	1.85	0.73-3.48	.0792
Cirrhosis	6	3.66	1.64	0.59-3.21	.1423
Kidney Cancer					
Melanoma	6	1.83	3.28	1.18-6.44	.0125
Lung Cancer	18	11.35	1.59	0.94-2.40	.0400
Uterine Cancer	5	2.10	2.38	0.75-4.93	.0582
Brain Cancer	8	3.41	2.35	1.00-4.26	.0244
Obesity	**5**	**1.15**	**4.35**	**1.37-8.99**	**.0099**
Parkinson's Disease	6	1.83	3.28	1.18-6.43	.0143
Mouth Cancers					
Prostate Cancer	11	6.37	1.73	0.86-2.90	.0559
Ovarian Cancer	6	2.73	2.20	0.79-4.31	.0548
Chronic Airway Obst.	7	3.96	1.77	0.70-3.32	.0951
Gallbladder Cancer					
Colon Cancer	11	6.25	1.76	0.87-2.95	.0516
Pancreatic Cancer	6	3.40	1.76	0.63-3.46	.1131
Alcohol Related	4	1.33	3.00	0.78-6.65	.0455
Rectal Cancer	4	1.30	3.07	0.80-6.81	.0427
Suicide					
Alcohol Related	**18**	**6.78**	**2.65**	**1.57-4.02**	**.0005**
Emphysema	29	14.85	1.95	1.31-2.73	.0011
Myeloid Leukemia					
Prostate Cancer	17	9.28	1.83	1.06-2.81	.0158
Lymphoma	7	4.68	1.49	0.59-2.81	.1685
Aneurysm					
Alcohol Related	10	4.60	2.18	1.04-3.73	.0207
Congestive Heart	13	8.58	1.52	0.80-2.45	.0885
Parkinson's Disease					
Prostate Cancer	20	11.86	1.69	1.03-2.51	.0197
Nephritis					
Cardiomyopathy	4	2.05	1.95	0.51-4.33	.1271
Renal Failure	4	2.33	1.72	0.45-3.82	.1710
Chronic Airway Obstruction					
Alcohol Related	14	7.12	1.97	1.07-3.13	.0158
Bronchitis	14	5.47	2.56	1.40-4.08	.0024
Emphysema	35	20.65	1.69	1.18-2.30	.0031
Lung Cancer	59	37.71	1.56	1.19-1.99	.0010

Table 23 (continued)

Cause of Death	Observed	Expected	Relative Risk	95% CI	P-value
Prostate Cancer					
Rectal Cancer	22	13.52	1.63	1.02-2.38	.0212
Mouth Cancer	14	9.10	1.54	0.84-2.45	.0735
Colon Cancer	**86**	**63.73**	**1.35**	**1.08-1.65**	**.0049**
Ovarian Cancer					
Breast Cancer	**27**	**18.74**	**1.44**	**0.95-2.04**	**.0409**
Prostate Cancer	**23**	**15.37**	**1.50**	**0.95-2.17**	**.0400**
Stomach Cancer	18	7.51	2.40	1.42-3.63	.0013
Uterine Cancer	9	3.52	2.55	1.16-4.49	.0122
Brain Cancer	10	5.75	1.74	0.83-2.99	.0630
Lung Cancer					
Colon Cancer	**79**	**53.08**	**1.49**	**1.18-1.83**	**.0007**
Mouth Cancer	11	7.47	1.47	0.73-2.47	.1210
Esophageal Cancer	10	6.29	1.59	0.76-2.73	.0951
Liver Cancer	12	7.32	1.64	0.84-2.70	.0643
Uterine Cancer	19	11.15	1.70	1.02-2.56	.0207
Emphysema	**54**	**33.11**	**1.63**	**1.22-2.09**	**.0007**
Stomach Cancer	34	22.91	1.48	1.03-2.02	.0183
Cirrhosis	32	21.16	1.51	1.03-2.08	.0375
Lymphoma					
Melanoma	8	4.21	1.90	0.81-3.44	.0606
Myeloid Leukemia	9	4.91	1.83	0.83-3.23	.0582
Cirrhosis					
Alcohol Related	**6**	**3.73**	**1.61**	**0.58-3.15**	**.0582**
Ulcer	10	4.83	2.07	0.99-3.55	.1515
Esophageal Cancer	**5**	**2.01**	**2.49**	**0.79-5.15**	**.0505**
Lung Cancer	36	19.84	1.81	1.27-2.46	.0010
Brain Cancer					
Colon Cancer	22	13.65	1.61	1.01-2.36	.0233
Stroke	45	29.00	1.55	1.13-2.04	.0040
Breast Cancer					
Ovarian Cancer	**30**	**19.18**	**1.56**	**1.05-2.17**	**.0139**
Gallbladder Cancer	15	5.29	2.84	1.58-4.45	.0008
Pancreatic Cancer	41	26.92	1.52	1.09-2.02	.0075
Emphysema					
Bronchitis	**14**	**5.57**	**2.51**	**1.37-4.00**	**.0029**
Chronic Airway Obst.	30	17.13	1.75	1.18-2.43	.0037
Lung Cancer	57	37.57	1.60	1.21-2.05	.0008
Myocardial Infarction					
Obesity	**51**	**37.68**	**1.35**	**1.01-1.75**	**.0003**
Heart Disease	**3,343**	**2,545.80**	**1.31**	**1.27-1.36**	**.0000**

Table 24 Rankings for GIF and Relative Risk for Death Certificates (ordered by sum of rankings for the four values)

Cause of Death	Relative Risk Ranking	GIF Value Ranking	R. Risk P-value Ranking	GIF P-value Ranking
Aneurysm	**9**	**6**	**2**	**3**
Kidney Cancer	**3**	**1**	**15**	**1**
Congenital Anomalies	**1**	**2**	**17**	**2**
Suicide	**7**	**15**	**1**	**5**
Chronic Airway	**13**	**8**	**3**	**4**
Alcohol Related	2	3	14	14
Prostate Cancer	14	10	4	6
Diabetes	18	17	5	8
Myeloma	6	4	13	18
Ovarian Cancer	**17**	**11**	**18**	**11**
Mouth Cancers	4	5	21	27
Myeloid Leukemia	8	12	19	19
Parkinson's Disease	11	9	23	17
Lung Cancer	**20**	**25**	**6**	**12**
Lymphoma	**22**	**16**	**20**	**7**
Motor Vehicle	**21**	**31**	**7**	**9**
Gallbladder Cancer	5	7	27	32
Emphysema	**29**	**14**	**22**	**16**
Myocardial Infarct.	32	32	9	10
Nephritis	12	20	24	28
Cardiomyopathy	**16**	**13**	**32**	**24**
Pulmonary Embolism	19	18	27	22
Stroke	35	29	11	13
Stomach Cancer	**28**	**19**	**29**	**15**
Brain Cancer	24	22	26	20
Breast Cancer	26	34	8	26
Circulatory Disorder	10	21	30	33
Colon Cancer	31	30	16	21
Hypertension	33	33	10	25
Cirrhosis	23	23	28	29
Alzheimer's Disease	15	24	31	34
Heart Valve Disorders	25	27	25	31
Congestive Heart	34	26	34	23
Heart Disease	36	38	12	39
Ulcer	30	28	35	35
Biliary, Gallbladder	27	36	33	37
Pneumonia	37	35	36	30
Pancreatic Cancer	39	37	38	36
Heart Conduction	38	39	37	38

to list where each cause of death ranked. The causes of death were ranked by raw GIF, GIF p-value, relative risk, and relative risk p-value. Only those causes of death that had both a relative risk and GIF were ranked. The sum of the rankings for the four parameters was used to order the causes of death in the table. Aneurysm was ranked the highest, followed by kidney cancer, congenital anomalies, suicide, and chronic airway obstruction. Kidney cancer and congenital anomalies would have ranked higher, but they had high p-values with the relative risk. This is most likely due to the small sample sizes of these groups.

Some causes of death such as lung cancer and motor vehicle accidents ranked higher in the relative risk results than the GIF results. This suggests a lack of genetic predisposition since the increased first-degree relative risk could be attributed to sharing the same environment or being involved in the same motor vehicle accident. Other causes of death such as ovarian cancer, lymphoma, emphysema, cardiomyopathy, and stomach cancer ranked higher in the GIF results, which suggests a genetic predisposition since the GIF examines both close and distant relationships.

4.7 Comparison of Death Certificate Results to Linked Cancer Records Results

The results of the GIF and relative risk for cancer deaths were compared to the results from the cancer registry papers (Cannon-Albright 1994; Goldgar 1994). The cancer registry records were linked to the UPDB genealogy records using probabilistic record linking which was discussed in Chapter 3. For most types of cancer, there were more cases in the cancer registry records than there were in the death certificates. For example, in the cancer records used for the cancer registry study, there were 8,060 prostrate cancer cases whereas there were 2,481 prostate cancer cases in the linked death certificates. The number of cancers ascertained by the death certificates links is incomplete, since many of the individuals in the cancer registry do not die from cancer. For example, there are 2,065 colon cancer cases from the cancer registry that are also in the death certificates. Only 55% of the individuals who were diagnosed with colon cancer actually died from colon cancer. Other discrepancies are possibly caused by the severity of the cancer and increased awareness in a family when a cancer death occurs. A cancer death in a family could lead to increased knowledge of the disease and preventative measures that would reduce the chances of additional cancer deaths. A study examining the value of routine screening in high-risk colon cancer families showed that the screening reduced the rate of colorectal cancer and seemed to prevent colorectal cancer deaths (Jarvinen 1995).

The GIF results for the death certificates and the cancer registry are compared in Table 25. The cancer registry results are taken from the cancer registry paper (Cannon-Albright 1994). The rankings for the death certificates and the cancer registry records are listed in Table 26 by raw GIF score and p-value. Both sets had similar rankings for myeloma, prostate cancer, brain cancer, rectal cancer, and pancreatic cancer. The two sets produced similar results when they

Table 25 GIF for Death Certificates Compared to Results for Cancer Registry Records

	Death Certificates			Cancer Registry		
Type of Cancer	N	GIF	P-value	N	GIF	P-value
Kidney	407	5.44	.0000	781	3.13	.067
Hodgkin's Disease	204	4.95	.0186	383	2.79	.393
Myeloma	**553**	**4.15**	**.0034**	**628**	**3.96**	**.0008**
Mouth Cancers[a]	277	4.12	.0728	825	4.75	.0000
Gallbladder	243	4.00	.1428	324	3.68	.055
Lymphoid Leukemia	**292**	**3.95**	**.0569**	**600**	**6.30**	**.000**
Prostate	**2,481**	**3.88**	**.0000**	**8,060**	**3.70**	**.000**
Ovarian	738	3.87	.0003	966	3.38	.0001
Myeloid Leukemia	461	3.86	.0058	629	3.39	.006
Lymphoma	969	3.67	.0000	1,986	3.38	.0001
Stomach	976	3.56	.0014	1,034	3.17	.006
Cervical	166	3.53	.3091	1,031	3.12	.022
Brain	**646**	**3.50**	**.0112**	**571**	**3.58**	**.004**
Lung	**2,120**	**3.28**	**.0003**	**2,477**	**3.33**	**.0001**
Colon	2,246	3.14	.0116	3,350	3.53	.0001
Liver	292	3.06	.4071	169	2.95	.368
Rectal	457	3.04	.3625	1,312	3.05	.044
Breast	2,203	3.03	.0450	5,811	3.23	.000
Melanoma	**382**	**2.99**	**.3703**	**1,157**	**4.06**	**.0001**
Pancreatic	**1,171**	**2.86**	**.4016**	**959**	**2.90**	**.268**
Uterine	416	2.86	.5760	1,945	2.90	.205
Bladder	501	2.62	.7359	1,837	2.90	.031
Connective Tissue	163	2.49	.6204	314	3.72	.002

[a] The GIF for lip cancer was calculated with cancer records.

Table 26 Rankings for GIF for Death Certificates Compared to Results for Cancer Registry Records

Type of Cancer	Death GIF	Cancer GIF	Death P-value	Cancer P-value
Kidney	1	16	1	19
Hodgkin's Disease	2	23	11	23
Myeloma	**3**	**4**	**7**	**9**
Mouth Cancers[a]	4	2	14	2
Gallbladder	5	7	15	18
Lymphoid Leukemia	6	1	13	1
Prostate	**7**	**6**	**2**	**3**
Ovarian	8	12	4	10
Myeloid Leukemia	9	10	8	14
Lymphoma	10	11	3	7
Stomach	11	15	6	13
Cervical	12	17	16	15
Brain	**13**	**8**	**9**	**12**
Lung	14	13	5	8
Colon	15	9	10	6
Liver	16	19	20	22
Rectal	**17**	**18**	**17**	**17**
Breast	18	14	12	4
Melanoma	19	3	18	5
Pancreatic	**20**	**22**	**19**	**21**
Uterine	21	21	21	20
Bladder	22	20	23	16
Connective Tissue	23	5	22	11

[a] The GIF for lip cancer was calculated with cancer records.

had the same sample size as in lung cancer or myeloma. Since both the death certificates and the cancer registry are linked to the genealogy database, the sets for some of the cancers most likely contain the same individuals. Large differences were observed for kidney cancer, hodgkin's disease, connective tissue cancer, melanoma, and lymphoid leukemia. The sample sizes for the death certificate sets for melanoma, connective tissue cancer, and lymphoid leukemia are much smaller than the cancer registry sets.

The mean of the control sets was higher for the death certificates sets. The average of the control means for the death certificates was 2.88, whereas the average for the cancer registry sets was 2.71. The age of the death certificate population for cancer deaths is likely higher than the cancer registry population. The older death cases would match to controls from cohorts where the genealogy database is more complete, which could produce the higher control means. The higher control values increase the p-values which affect the rankings by p-value and make it more difficult to compare the results of the two data sets.

The rankings for the cancer's relative risk are listed in Table 27. The cancer registry values come from the cancer registry relative risk paper (Goldgar 1994). There were no p-values published with the paper, so a ranking by p-value was not done. As with the GIF results, where there were large sample sizes for the death certificates, the relative risk values were close. This was the case for breast cancer, stomach cancer, and brain cancer. The rankings were similar for both sets, except for colon cancer. This difference can likely be attributed to the fact that

Table 27 Rankings for Relative Risk for Death Certificates Compared to Results for Cancer Registry Records

	Death Certificates			Cancer Registry		
Type of Cancer	N	Relative Risk	Ranking	N	Relative Risk	Ranking
Kidney	391	6.08	1	687	2.45	4
Gallbladder	228	5.89	2	253	2.13	6
Myeloid Leukemia	445	4.01	3	749	2.97	1
Prostate	2,259	2.82	4	6,350	2.21	5
Ovarian	715	2.57	5	883	2.05	8
Lung	2,046	2.19	6	2,228	2.55	3
Lymphoma	928	2.15	7	1.362	1.68	11
Brain	**584**	**2.00**	**8**	**1,220**	**1.96**	**9**
Breast	**2,084**	**1.88**	**9**	**5,559**	**1.83**	**10**
Stomach	937	1.73	10	800	2.09	7
Colon	**2.085**	**1.53**	**11**	**2,861**	**2.67**	**2**
Bladder	460	1.19	12	1,452	1.53	12
Pancreatic	1,104	0.90	13	749	1.25	13

many individuals with colon cancer had another cause of death.

4.8 Cause of Death Discussion

The results of the GIF analysis and the familial relative risk analysis for each cause of death were examined. Also studied were medical, environmental, and genetic references for each cause of death. A brief summary of the risk factors for each disease is given. The main references used are a medical textbook (Andreoli 1993) and two medical genetics books (Weatherall 1991; Jorde 1995). Another source which provided a great deal of information was the Online Mendelian Inheritance of Man (OMIM) at Johns Hopkins University. OMIM contains genetic references for any disease or disorder that has been associated with a genetic cause. It is accessible through the internet. The order of the causes of death for this review comes from the original list of causes of death in Table 7. The p-values are listed with the GIF scores and relative risk scores when they are relevant to the discussion.

Breast cancer is the most common malignancy of women in the United States. It was formerly the leading cause of cancer death among women but has recently been surpassed by lung cancer. In the linked Utah death certificates, it was the most common cause of cancer deaths for women. It has a number of risk factors such as increasing age, first- and second-degree relatives with breast cancer, age at first pregnancy, early menstruation, and radiation therapy to the chest. The role of genetic predisposition has been strongly identified in breast cancer cases. If a woman has one first-degree relative with breast cancer, her risk of developing the disease doubles. The risk is higher if the onset of disease is at an early age and it is bilateral (Ottman 1983).

An autosomal dominant gene BRCA1, which is believed to account for approximately 5% of the breast cancer cases in the United States, has been cloned (Miki 1994). It has been shown that the penetrance of the gene is 0.92 by age 70 (Goldgar 1992). Mutations in the BRCA1 gene have been found in families with excess breast cancer and ovarian cancer. It has also been suggested that male carriers of the mutated gene have an increased risk of prostate cancer (Arason 1993). There is also another gene linked to breast cancer that is known as BRCA2 that was cloned in 1995 (Wooster 1995). It has been identified in studies of breast cancer families where there were male breast cancer cases. A study with 145 breast-ovarian families showed that 76% of the families were linked to BRCA1. However 13 of the families that had male breast cancer cases in them did not link to BRCA1 (Narod 1995). It is thought that these families could be linked to BRCA2. Other types of cancer that have been associated with breast cancer include colon cancer, stomach cancer, and pancreatic cancer (Lynch 1987).

The GIF score for breast cancer was only 3.03 (P = .045) with the control 2.81. The average age of death is 65.51, so it is possible that many of the breast cancer deaths are sporadic cases and not familial. The GIF for the youngest third (age of death under 60) is 3.15 (P = .1178) with the control equal to 2.81. However the GIF for the 315 breast cancer cases where the age of death was 50 or less was 4.84 with a control of 3.05. This shows a stronger familial component in cases where the age of death is young.

The first-degree relative risk for breast cancer was 1.88 (P = 0). For the youngest (under age 60) set of cases, it was 2.11 (P = .0029). The risk of ovarian cancer in families where the probands are breast cancer deaths was 1.56. Other cancers that showed increased risk with breast cancer probands are gallbladder cancer at 2.84, uterine cancer at 1.57, pancreatic cancer at 1.51, and prostate cancer at 1.27. There was an increase in the familiality for breast cancer that was strongest in the younger cases. Also an increased risk for ovarian cancer that is consistent with the research done on the BRCA1 gene was seen.

The cause of **colon cancer** and **rectal cancer** is unknown, although a number of risk factors have been identified. Environmental factors, particularly diet have been implicated most. A diet low in fiber and high in animal fat and protein has been suggested as an important risk factor. This type of diet correlates to the regions of the world where the incidence of colon cancer is high. There is also evidence that hereditary plays a role in colon cancer.

Studies have shown clustering of colon cancer in families. This research has led to the discovery of four genes responsible for hereditary nonpolyposis colorectal cancer (HNPCC), which may account for 10% of all colorectal cancer cases (Froggatt 1995). This disorder is characterized by a dominantly inherited predisposition to early onset colon cancer. There is also evidence that colon cancer can develop from benign adenomatous polyps. A gene known as APC has been discovered that is linked to the development of polyps, which often change into malignancies. The trait is known as familial adenomatous polyposis. It affects about 1 in 8000 individuals. HNPCC accounts for a much larger proportion of the hereditary colon cancer cases.

The GIF for colon cancer was 3.14 (P = .0116). The younger (under age 66) set had a score of 3.43 (P=.0492), and there was no difference between males and females. The relative risk of colon cancer was 1.56 in all the cases and 1.83 in the younger cases. The risk for males was 1.90 compared to 1.18 for females. There was also an increased risk for kidney cancer (risk = 1.49), and uterine cancer (risk = 1.29) for relatives of the colon cancer probands. These cancers have been found in HNPCC families. There is a familial aggregation for colon cancer, especially in the younger cases.

Prostate cancer is the most common cancer of men and the second most common cause of cancer deaths in the United States. It was the most common cause of cancer deaths in the linked Utah death certificates. It is rare in men under the age of 50, but the incidence increase steadily with age. Since prostate cancer usually occurs at an advanced age, patients often die as a result of other causes.

Studies have shown that there is a significant familial factor in the development of prostate cancer (Cannon 1982; Meikle 1985). An increased risk of prostate cancer was found for men with a brother or father affected (Steinberg 1990). Familial influence was also seen in young onset cases (Carter 1992).

The average age of death for prostate cancer in the linked death certificates was 77.47, so there are likely many men with prostate cancer who die from another cause of death. The GIF and relative risk results show a strong familial aggregation for prostate cancer, especially in the younger cases. The GIF for all of the cases was 3.88 (P = 0). The GIF for the younger cases (under age 73) was 4.61 (P = 0). The relative risk was 2.82 (P = 0) for all the cases and 4.25 for the younger cases (P = 0).

The risk factors for **ovarian cancer** are similar to breast cancer. Ovarian cancer has been linked to the BRCA1 gene, along with breast cancer (Narrod 1995). Mutations in BRCA1 have also been found in some sporadic cases of ovarian cancer (Merajver 1995). Women who live in industrialized countries are at a higher risk, but the use of oral contraceptives appears to decrease the risk. The diagnosis of ovarian cancer is frequently delayed because the symptoms are nonspecific. The majority of women who have ovarian cancer are diagnosed when the disease is advanced.

The GIF for all of the ovarian cancer deaths was 3.87 (P = .0003), and the GIF for the youngest (under age 60) set was 6.19 (P = 0). The relative risk for all cases was 2.57 (P = 0.143), and for the younger cases it was 2.36 (P = .1611). These scores show a strong familial influence that was not seen in the breast cancer scores. It is likely that a death in a family from breast cancer or ovarian cancer would lead to preventative measures and earlier diagnoses. Perhaps the severity of the ovarian cancer when it is diagnosed is so great that the increased awareness of family history does not have the same effect as it does with breast cancer.

The incidence rate of **melanoma** has risen higher than any other cancer except lung cancer. The main risk factor is sun exposure. Utah has a higher incidence rate, which is likely due to increased sun exposure. The disease is rarely fatal if detected and treated early.

A number of studies have shown a positive family history for melanoma that was correlated with early age of onset and a tendency for multiple primary lesions (Kopf 1986; Anderson 1967). It is believed that hereditary cases make up about 10% of all melanomas. The linkage of a melanoma susceptibility gene on chromosome 9 was reported in 1992 (Cannon-Albright 1992). Genomic clones, which were thought to be involved in susceptibility to melanoma and to influence progression of certain other tumors, were discovered in a region of chromosome 9 (Weaver-Feldhaus 1994). It is thought that this gene acts as a tumor suppressor. Not all of the families in these studies showed linkage to chromosome 9, so it is possible that there are other genes involved in hereditary melanoma. There have been studies that mentioned three different locations for linkage to melanoma, including chromosome 9 (Bergman 1994).

There were only 382 cases of melanoma in the linked death certificates. The GIF for all the cases was 2.99 (P = .3703), and the younger (under age 60) set GIF was 1.68 (P = .8301). There were not enough first-degree relative cases to calculate a relative risk. The GIF scores were much lower than those from the linked cancer records. Since melanoma is not fatal when diagnosed early, increased awareness in families could lead to reduced risk of dying from melanoma, thus making most of the melanoma deaths sporadic.

Chronic **lymphoid leukemia** is a disease of older persons, with fewer than 10% of cases where the patient is less than 50 years old. It affects twice as many males as females. Acute lymphoid leukemia affects primarily children and is more life-threatening than chronic lymphoid leukemia. Known risk factors associated with the development of acute lymphoid leukemia include radiation, viruses, genetic predisposition and chemicals. A study has reported familial aggregation of chronic lymphocytic leukemia and autoimmune disease (Fraumeni 1969). Another study found the same relationship (Conley 1980). They concluded that genetic factors in these families disturb the regulation of the immune system. A gene known as MLL (mixed lineage leukemia) has been associated with both lymphoid leukemia and myeloid leukemia (Ziemin-van der Poel 1991). Although lymphoid leukemia can be fatal, aggressive treatments with chemotherapy and bone marrow transplantation have been successful.

Chronic **myeloid leukemia** is a genetic disorder, specifically a somatic cell disorder. It does have environmental causes such as radiation and chemical (i.e., benzene) exposure. How these agents interact with bone marrow cells to produce a malignant clone that lacks the ability to differentiate into normal mature blood cells is not known. Myeloid leukemia affects primarily adults. A translocation involving chromosomes 9 and 22 seems to be the oncogenetic mechanism for myeloid leukemia. Two genes on these chromosomes are involved in the translocation, the BCR and ABL genes (Chissoe 1995). This translocation is associated with more than 90% of chronic myeloid leukemia, 25 - 30% of acute lymphoblastic leukemia, and 2 - 10% of childhood acute myelogenous leukemia.

There were 292 deaths from lymphoid leukemia and 461 deaths from myeloid leukemia. The GIF for lymphoid leukemia was 3.95 (P = .0569), and 3.06 for the younger (under age 70) cases. There were almost twice as many male deaths. The GIF for males was 3.73 compared to 2.29 for females. There were not enough first-degree relative cases to calculate a relative risk. The GIF scores do show a familial influence, especially in males.

The GIF for myeloid leukemia was 3.86 (P = .0058), and it was 3.31 (P = .2225) for the younger (under age 65) cases. The GIF for males was 4.98 (P = .0011) compared to 3.89 for females. The relative risk for all the cases was 4.01 (P = .0170) and was 4.40 (P = .0495) for males. These scores support the belief that myeloid leukemia is genetically predisposed.

The **mouth cancer** group has a number of different cancers that are located in the mouth area. Since death from most of these cancers was rare, they were grouped together. For some of them, such as lip cancer, it would be impossible to study them separately, since death from lip cancer was extremely rare. Risk factors for these cancers are tobacco and alcohol use. There is not a lot of genetic information on mouth cancers, although the two papers from which the GIF and relative risk methods were taken, showed that lip and oral cavity cancers had high degrees of familiality (Cannon-Albright 1994; Goldgar 1994).

The data from these studies do come from the same population as the death certificates, so similar results were expected. The GIF for all the cases was 4.12 (P = .0728) and was 5.52 for the younger (under age 70) set (P = .0142). There were not enough first-degree relative cases to generate a relative risk.

The cause of **esophageal cancer** is not known. Environmental factors are usually implicated, particularly in those areas of the world having the highest incidence. In the United States, tobacco use and alcohol abuse are considered primary risk factors for esophageal cancer.

There were only 228 cases of esophageal cancer. The GIF scores do not show any familial influence. The GIF for all the cases was 2.56 which was less than the control average. There were not enough first-degree relatives with esophageal cancer to calculate the relative risk.

Environmental factors have long been suggested as the cause of **stomach cancer**. The reasons include the high incidence of stomach cancer in specific regions of the world, particularly in Japan, and changes in incidence rates in migrating populations. A diet high in salt and nitrates is thought to be a potential environmental factor.

There are not any significant references on a genetic predisposition to stomach cancer. The relative risk from all of the death records was 1.73 (P = .0384) and the younger (under age 67) set had a relative risk of 2.00 (P = .1379). The GIF for stomach cancer in the linked death certificates was 3.56 (P = .0014), and it was 4.45 (P = .0031) for younger cases. It is possible that the high GIF scores are a result of families sharing the same diet, although there may be some kind of genetic susceptibility since there are higher scores for the younger cases.

The majority of **liver cancer** cases in the United States are due to metastases from other sites such as stomach, pancreas, colon, lung, bladder and from melanoma. Cancers where the liver is the primary site are rare in the United States. In other parts of the world such as sub-Sahara Africa, China, Japan and southeast Asia, it is one of the most frequent malignancies. Liver cancer often arises in a cirrhotic liver and in closely associated with chronic hepatitis B or C virus infection. The advent and widespread use of vaccinations to prevent infection with hepatitis B virus are expected to reduce the incidence of liver cancer. It is the only disease for which immunization against a malignancy is currently available. This explains the low incidence rate in the United States and the higher rates in other parts of the world where the vaccine is not available.

The risk of liver cancer is intermediate in cirrhosis due to alcohol and high in hemochromatosis. The term hemochromatosis refers to an increase in total body iron stores with iron deposition in parenchymal tissues that leads to functional impairment of the most severely affected organs. The liver is usually the first organ that is affected. Liver cancer develops as a sequel to the cirrhosis caused by hemochromatosis in about 35% of the cases.

Hemochromatosis occurs both sporadically and in families. Familial hemochromatosis is linked to the human leukocyte antigen (HLA) locus

(Edwards 1980) and is clinically manifest in roughly 1 in 5000 Caucasians in the United States. It is inherited as an autosomal recessive trait. Homozygotes have large iron stores, but only a minority of them manifest the disease. It is observed 5 to 10 times more commonly in men than women.

The number of liver cancer deaths was only 292. The GIF score for all the cases was 3.06 which was not significantly different from the controls, but the GIF for the younger (under age 65) cases was 6.11 (P = .0125) and the GIF for the men was 4.83 (P = .0688). This could be a result of the risk from alcoholism and hemochromatosis, since both occur more frequently in males and are familial. However, there was not a significant association between liver cancer, alcohol related deaths, and cirrhosis. There were not enough first-degree relatives to calculate a relative risk with liver cancer as the primary site, but the relative risk of liver cancer was high where breast cancer and lung cancer were the primary sites. These cases could be metastatic liver carcinomas.

Gallbladder cancer is rare. Symptoms resemble cholecystitis or bile duct obstruction. It is often diagnosed when it is advanced, and the prognosis for survival is poor. The GIF score for the death cases was 4.00 (P = .1428). The score for men in the death certificates was 6.34 (P = .0233). Even though, there were only 243 cases, there were enough first-degree relatives to calculate a relative risk of 5.89 (P = .0485). These numbers suggest a familial influence. The causes of death that have high relative risks when gallbladder is the primary site include colon cancer, rectal cancer, pancreatic cancer, and alcohol-related deaths.

Pancreatic cancer is an almost uniformly fatal malignancy. It is the fourth most common malignant tumor, accounting for 5% of cancer deaths in the United States. The cause of pancreatic cancer is unknown. Studies have identified risk factors such as advanced age, smoking, diabetes, some forms of chronic pancreatis, and dietary habits such as increased consumption of animal fat and protein. Somatic mutations of the *p53* proto-oncogene have been found in pancreatic tumors (Casey 1993). *p53* is medically important, since the presence of *p53* mutations signal a more aggressive cancer with relatively poor survival prospects.

The GIF and relative risk results show no familial influence for pancreatic cancer. The GIF for all death cases was 2.86 (P = .4016), and the relative risk was less than one.

Lung cancer is the leading cause of cancer deaths in the United States. Cigarette smoking is the most important risk factor. Lung cancer is 10 to 30 times more common among smokers. Approximately 4% of those who have smoked for more than 40 years develop lung cancer. The rate of lung cancer is lower in Utah than the national rate. Utah has the lowest smoking rate of any state in national surveys. This can be attributed to the teachings of the LDS church which prohibit the use of tobacco and alcohol.

The biology of lung cancer has received a lot of attention, particularly regarding the role of oncogenes and other genetic mechanisms of tumor development. Mutations of the *p53* gene have been found in 50% of lung cancers.

Activation of the KRAS oncogenes have also been found in lung cancer cell lines (Nakano 1984). Although the role of oncogenes in the development of lung cancer is unknown, their expression has been associated with decreased survival. Other genetic factors associated with lung cancer include the variable expression of certain cytochrome P-450 enzymes. This enzymatic activity is inducible by cigarette smoke, so one factor possibly contributing to lung cancer may be the genetically regulated activity of these or related enzymes.

The GIF score of lung cancer for all cases was 3.28 (P = .0003) and was 3.47 (P = .0567) for the younger (under age 64) set. It was much higher in women than in men. The female score was 4.13 (P = .0084) and the male score was 3.28 (P = .0040). There were four times as many male deaths as female deaths. The relative risk for all cases was 2.19 (P = 0). The relative risk values for males and the younger set were similar to the score from all the cases, whereas the score for women was 2.68 (P = .0005). The results show that there is a familial predisposition for lung cancer that is especially strong in first-degree relatives.

There were only 163 deaths from **connective tissue cancer**. The GIF score for all cases was lower than the control. The GIF scores for the younger set and the male set were high, but since there is such a small sample set, their validity is questionable. Connective tissue cancer did have an elevated risk in relatives of probands with both types of leukemia.

Uterine cancer occurs most often in postmenopausal women. Risk factors include obesity, previous pelvic radiation therapy, and estrogen replacement therapy. The GIF for all the cases was 2.86, which was similar to the controls. The score for the younger (under age 70) cases was slightly higher at 3.32 (P = .4494), but there was not a significant difference between the cases and the controls. There does not appear to be any familial influence.

Cervical cancer accounts for 2.5% of all the malignancies of women in the United States. Since the advent of cervical and vaginal cytology in the early 1940s, the incidence and mortality of cervical cancer have been decreasing. The major cause of cervical cancer is a genital human papilloma virus (HPV). The DNA of HPV types 16 and 18 has been found closely associated with human genital cancers, supporting an etiologic role for these viruses (Durst 1987).

The GIF for all cervical cancer deaths was 3.53 (P = .3091), but there was not a significant difference between the controls and the cases. There were only 166 cervical cancer deaths. There were not enough first-degree relative cases to calculate the relative risk.

Bladder cancer accounts for 3% of all the malignancies in the world. It is three times more common in males than in females. There is substantial geographical variation in the incidence of bladder cancer. Incidence rates are higher in white male populations in developed countries. Risk factors include smoking and exposure to carcinogenic chemicals. A study has demonstrated that there are genetic susceptibility factors in smoking and occupational-related bladder cancers (Risch 1995). Deletions involving chromosome 9 represent the most frequent genetic change identified in bladder tumors. It was of particular interest that these deletions were present at similar frequency in bladder tumors at all grades and stages (Tsai 1990). This finding of chromosome 9 deletions as the sole genetic change in many low-grade, early-stage tumors suggests that it may represent an early or initiating genetic event (Keen 1994). The possible familial components for bladder cancer could be exposure to the same carcinogenic agent and a genetic susceptibility to that agent.

The GIF scores for all the bladder cancer deaths do not show any familial component, but the younger (under age 70) set has a GIF of 4.03 (P = .2334). This is high, but not significant since there are not many cases. However some of the younger set could have the genetic susceptibility discussed above. There were not enough first-degree relatives with bladder cancer to calculate a relative risk.

Kidney cancer accounts for about 1.5% of all the cancer cases in the world. Tobacco is a well established risk factor, along with lesser risks of industrial exposure to airborne aromatic hydrocarbons from coke production and the abuse of analgesics containing phenacetin.

A number of studies have shown that some kidney cancers are inherited. One study described a family in which members with an inherited chromosomal translocation were predisposed to renal cancer (Cohen 1979). Another study reviewed nine families in which two or more members had kidney cancer (Li 1982). Multiple generations were affected in five families and siblings were affected on the other four families. The median age of diagnosis was a decade earlier than average. Individual patients had bilateral or multifocal lesions, which are features of hereditary forms of diverse cancers. None of the patients had the translocation described in the other study. A more recent study examined 28 families with multiple cases of renal cancer (Levinson 1990).

Kidney cancer had the second highest GIF score of all the causes of death with 5.44 (P = 0). The GIF for the younger (under age 65) set was 7.21 (P = .0009) and the female GIF was 7.20 (P = .0002). A characteristic of inherited kidney cancer is early age of onset. The relative risk for all the kidney cancer deaths was 6.08 (P= .0012), and the relative risk for the younger set was 6.17 (P = .0455). The relative risk for kidney cancer in the linked cancer records was 2.45. There is definitely a familial predisposition for kidney cancer that is shown by the results.

Brain cancer can arise anywhere in the intracranial cavity. Most intracranial tumors begin in the brain, but they may be the site of a metastatic spread from tumors that arise outside the nervous system. Metastatic intracranial tumors are equal to or greater in number than primary neoplasms. Some cancers that often metastasize to the brain include lung, breast, and melanoma. One risk factor seems to be advanced age, since the incidence of brain cancer is rising as the population ages. Hereditary factors would include the genetic predisposition of the cancers that metastasize to the brain.

The GIF for all brain cancer deaths was 3.50 (P = .0112). The GIF for the younger cases was not significant, but the GIF for females was 5.00 (.0003). This could be a result of metastasizing breast cancer. The relative risk for all cases was 2.00 (P = .0384), and the highest relative risk was for females at 2.61 (P = .0934).

These results show a familial influence, especially in women.

Hodgkins disease is a disease of young adults, but it does occur in children and the elderly. Risk factors are environment and hereditary. The cause of the disease seems to differ in the old and the young cases (MacMahon 1966). Relatives of young adults with Hodgkins disease are at increased risk. One study concluded that genetic susceptibility underlies Hodgkins disease in young adulthood (Mack 1995). A link between Hodgkins disease and Epstein Barr virus was suggested in epidemiological studies (Munoz 1978; Mueller 1989).

The GIF score for all Hodgkins disease deaths was 4.95 (P = .0186) and for the younger (under age 55) set it was 5.29 (P = .0844). It was especially high in females at 10.38 (P = .0004), but there were only 68 female cases. The high score in the younger set would support the belief that the younger onset cases are familial. There were not enough first-degree relative cases to calculate a relative risk.

Malignant **lymphomas** other than Hodgkins disease are a heterogeneous group comprised of Burkitt's lymphoma, lymphosarcoma, and reticulosarcoma. Burkitt's lymphoma is a rare neoplasm that has been causally related to the Epstein-Barr virus. The other type of lymphomas are classified by cell type and clinical stage of the disease. As in Hodgkins disease, the cause of lymphoma is not known. Viruses, radiation, immunosuppression (i.e., organ transplantation, AIDS) and certain genetic conditions have been implicated.

Familial lymphoma is uncommon, and it is usually associated with various forms of immunodeficiencies. A study described two sisters in an American family who died of Burkitt's lymphoma at ages 11 and 22 years (Anderson 1986). The mother and two healthy brothers had abnormality of lymphocyte subsets. An inherited disturbance of lymphocytes was thought to underlie the familial aggregation for Burkitt's lymphoma.

The GIF for all the lymphoma cases was 3.67 (P = 0) and for the younger (under age 65) cases it was 3.92 (P = .0461). The relative risk for all the cases was 2.15 (P = .0174), and for the younger set it was 3.25 (P = .0234). These numbers show that familial lymphoma does exist, and it is stronger in younger cases.

Myeloma is a malignant disease of plasma cells that is characterized by the presence of monoclonal immunoglobin or light chains in the serum and urine and bone destruction. The typical patient is over 50, and it occurs two times more frequently in blacks than whites in the United States. Ionizing radiation can cause myeloma, and increased risks of the disease have been observed among survivors of the atomic bomb explosions at Hiroshima and Nagasaki, among women given radiation treatments for cervical cancer, and among workers in the nuclear industry.

Most family studies of myeloma focus on immunological disorders. One study described 19 cases of familial immunopathy, distributed in nine families (Zawadzki 1977). Ten members of five families had multiple myeloma, five members of two families had lanthanic paraproteinemia and four members of two families had one or the other of these disorders. Two studies reported that identical twins were concordant for myeloma (Comotti 1987; Judson 1985).

The GIF for all the myeloma deaths was 4.15 (P = .0034), and for the younger (under age 70) cases it was 5.20 (P = .0006). The relative risk for all the myeloma deaths was 4.78 (P = .0008), and the relative risk for the younger cases was 7.01 (P = .0011). This numbers show a strong familial influence, that could be a result of immunological genetic disorders or environmental exposure of families to radiation.

Diabetes mellitus is a very common disorder, with an estimated prevalence of 2 - 4% in the United States. The complications of diabetes account for more than 25% of all end stage renal failures and more than 50% of lower extremity amputations. It is also the leading cause of blindness. There are two major types of diabetes: type 1 (insulin dependent diabetes mellitus, IDDM) and type II (non-insulin-dependent diabetes mellitus, NIDDM).

The peak age of onset for IDDM is between 11 and 13 years, coinciding with the onset of puberty, but the IDDM can begin at any age, including in the elderly. The etiology of IDDM is unknown. A leading hypothesis is that a viral illness or another unspecified initiating event may damage the beta cells of the pancreas, followed by a slow autoimmune destruction of the remaining beta cells in susceptible individuals. This autoimmune hypothesis also accounts for the increased risk of developing diabetes in individuals with certain HLA genes. The genes that control the autoimmune response are located on the sixth chromosome close to the HLA loci (Donald 1989).

NIDDM is much more common than IDDM with approximately 10 cases of NIDDM for each case of IDDM. It usually has its onset after age 40. The two most important risk factors are obesity and family history. Identical twins are almost 100% concordant for NIDDM, suggesting a very strong genetic component for this disorder. Despite the apparent high degree of genetic involvement in NIDDM, specific genes for this disorder have not been identified.

The GIF for all cases of diabetes was 3.64, and the younger (under age 70) set was 4.36. The p-values for all of the calculations for diabetes was zero, due to the large number of cases and the strong familial tendency. The relative risk for all cases was 2.38, and for the younger cases it was 3.49. These values demonstrate a strong familial predisposition for diabetes.

Obesity is largely genetically determined. A child of two obese parents has about an 80% chance of becoming obese, whereas the risk is only 15% for the offspring of two parents of normal weight. Several studies have predicted the presence of an obesity gene (Paganini-Hill 1981; Zonta 1987). A mouse obesity gene was cloned by positional cloning (Zhang 1994). The obesity gene product is present as a 16-kD protein in mouse and human plasma (Pelleymounter 1995). Data from this study suggested that the obesity protein regulates body weight and fat deposition through effects on metabolism and appetite. Another study suggested that the obesity protein serves an endocrine function to regulate body fat stores (Halaas 1995).

The GIF for obesity was 3.34 (P = .3363) for all cases. It was 4.00 (P =

.0757) for the female cases. There were only 226 cases in which obesity was listed as a cause of death. The majority of the individuals who would be classified as obese would likely die from some other disease. This makes it difficult to do a complete analysis due to the small number of cases in the death certificates.

Senility with psychosis is characterized by patients who become so apathetic as to seem depressed. They suffer great anxiety, increased irritability, paranoia, or secondary depression. In contrast to Alzheimer's and other progressive dementias, a lack of social amenities characterizes the mental deterioration that accompanies frontal lobe disease, intracranial mass lesions, or chronic drug-alcohol abuse. Aging persons are especially susceptible to chronic drug intoxication and depressive illness. The results of the GIF did not show any familial tendency for this cause of death.

Senility without psychosis was a general diagnosis that possibly included many patients with **Alzheimer's disease**. Since there is a large amount of information on Alzheimer's disease the focus of the summary on these two causes of death will be on Alzheimer's.

It is estimated that Alzheimer's disease affects approximately 10% of Americans over the age of 65 and up to half of those over 85. The disorder is characterized by progressive dementia, loss of memory, and the formation of amyloid plaques and neurofibrillary triangles in the brain. Death usually occurs 5 to 10 years after the first appearance of symptoms. It is a difficult disease to diagnose, since a definitive diagnosis can only be obtained by a brain autopsy.

The cause of Alzheimer's disease is not known, but recent attention has focused on a possible hereditary factor associated with an abnormality on chromosome 21. This seems to be especially true in families in which there is early onset of the disease (Goate 1991). At-risk individuals in early offspring families had an estimated lifetime risk for dementia of 53% (Farrar 1990). The researchers speculated that this was a result of autosomal dominant inheritance. The lifetime risk in late-onset families was 86%. The researchers concluded that this form may have at least two causes: autosomal dominant inheritance in some families and other genetic or shared environmental factors in other families. The difficulties in diagnosing Alzheimer's disease hinders its genetic analysis. Also, since the age of onset can be very late, individuals carrying the gene for Alzheimer's could die from another cause before developing the disease.

The GIF for all the deaths from senility without psychosis was 3.69 (P = .0265), and the GIF for Alzheimer's disease was 3.31 (P = .1907). The GIF for the two causes of death combined was 3.32 (P = .0072). The GIF for Alzheimer's disease for cases under age 80 was 4.89 (P = .0064), which shows the familial aggregation of younger onset cases. The relative risk for Alzheimer's was 2.80 (P = .0823) for all the cases, and for the younger cases, the relative risk was 3.24 (P = .1038). The average age of death for both causes of death was more than 81.

Alcohol-related deaths include death from acute alcoholic intoxication, alcohol psychosis, and alcohol abuse. At some point in their lives, alcoholism is diagnosed in approximately 10% of males and in 3 to 5% of females (Jorde 1995).

Twin and adoption studies show that alcoholism clusters strongly in families. This reflects a possible genetic contribution to the disease. One study identified two separate heritable types of alcoholism (Cloninger 1987). Type I alcohol abuse has its usual onset after the age of 25 years and is characterized by severe psychological dependence and guilt. It occurs in both men and women and requires both genetic and environmental factors to become manifest. By contrast, type II alcohol abuse has its onset before the age of 25. Persons with this type of alcoholism are characterized by their inability to abstain from alcohol and by frequent aggressive and antisocial behavior. Type II alcoholism is rarely found in women and is much more heritable than type I.

There were 399 alcohol-related deaths, with 324 of them male. The average age of death was 60.72, which was low when compared to most of the other causes of death. The GIF for all the alcohol-related deaths was 4.34 (P = .0005) and it was 5.44 (P = 0) for males. The relative risk for all cases was 6.13 (P = .0012) and it was 7.46 (P = .0006) for males. These results show a strong familial predisposition for alcoholism, especially in males.

Parkinson's disease, an idiopathic disorder of adults, has its highest incidence in men over 40 years of age. Epidemiologic studies have traced some cohorts to long-preceding influenza epidemics. One study postulated that Parkinson's disease is the result of environmental factors acting on genetically susceptible persons (Barbeau 1985). Another study found that the cumulative risk of the disease among siblings of probands with affected parents was significantly higher over that for siblings of probands without affected parents (Lazzarini 1994).

The GIF results and the relative risk scores do show a familial effect for Parkinson's disease. The GIF for all deaths was 3.91 (P = .0021), and for the younger (under age 75) cases it was 5.75 (P = .0076). It was also high in women at 5.18 (P = .0253). The relative risk for all the deaths was 3.45 (P = .0250), and for the younger cases it was 4.00 (P = .0793).

The **motor neuron disease** group includes amyotrophic lateral sclerosis, progressive muscular atrophy and bulbar palsy. The majority of the deaths in this group were from amyotrophic lateral sclerosis. Amyotrophic lateral sclerosis usually is sporadic, but familial groupings have occurred, indicating a genetic predisposition or common exposure to an unknown causative agent. Familial cases tend to affect younger persons and to progress more rapidly then do sporadic ones.

About 10% of amyotrophic lateral sclerosis (ALS) cases are familial (Pramatarova 1995). Tight genetic linkage between ALS and the gene for Cu/Zn-binding superoxide dismutase (SOD1) was reported (Rosen 1993). One study demonstrated that mutation in the SOD1 gene can also be responsible for sporadic cases of ALS (Jones 1993).

There were only 202 deaths from motor neuron diseases. The GIF for all cases was 4.16 (P = .0647), and it was 14.75 (P = 0) in the younger (under age 65) cases. These results show a familial influence, although there are not many cases.

There were not enough cases to calculate a first-degree relative risk.

Multiple sclerosis is the most common of the presumed immune demyelinating disorders of the central nervous system. It usually causes its first symptoms between the ages of 20 and 40 years and is characterized by remissions and exacerbations of neurologic dysfunction affecting several different sites in the central nervous system over many years. Statistically, the disorder does not greatly decrease life expectancy, although some middle-aged patients become severely disabled and die prematurely of complications.

The etiology of multiple sclerosis is unknown, although most clues indicate immunologic and genetic factors. Genetic predisposition is suggested by the strong association with the haplotype HLA-DW2, which indicates an immune-response mechanism (Terasaki 1976). Another study confirmed that a MS genetic susceptibility gene exists in the HLA complex (Francis 1987). A recent study concluded that familial aggregation in MS is genetically determined (Ebers 1995).

Multiple sclerosis had the highest raw GIF score for all the deaths. It was 6.86 (P = .0002). The younger (under age 55) set GIF was 6.25 (P = .0257), and the GIF for women was 9.26 (P = 0), which was substantially below the control score in men. These findings are similar to the recent research. There were not enough first-degree relatives with MS for a relative risk calculation, since there were a small number of deaths from MS.

Heart valve disorders include diseases and disorders of the mitral, aortic, and tricuspid valves. The most common cause of mitral stenosis is rheumatic fever. Mitral valve prolapse is common in Marfan's syndrome and other connective tissue diseases. It is also more common in females than males and occasionally will run in families. Familial occurrence of mitral valve prolapse was reported in several studies (Hunt 1969; Shell 1969). Aortic valve disorders are usually congenital or caused by rheumatic fever. Disorders with the tricuspid valve are also linked to rheumatic fever. Rheumatic fever usually occurs in children and is caused by group A betahemolytic streptococcal pharyngitis.

The GIF scores do not show a significant familial influence for heart valve disorders, but the relative risk does. The relative risk for all cases is 1.97 (P = .0314), and it is 2.50 (P = .0359) in men. This could be attributed to familial cases of rheumatic fever.

Essential **hypertension** is a common disease believed to result from the interplay of multiple genetic and environmental determinants. It is a key risk factor for heart disease, stroke, and kidney disease. The most important environmental risk factors for hypertension are increased sodium intake, decreased physical activity, stress, and obesity, which as discussed earlier is influenced by genetic factors.

Blood pressure regulation is a complex process that is influenced by many physiologic systems. These include various aspects of kidney function, cellular ion transport, and heart function. Most research for hypertension is focused on specific components that may influence blood pressure variations such as angiotensin, angiotensinogen, and sodium-lithium transport.

A number of studies demonstrate impaired sodium transport in various ways (Garay 1980; Weder 1986). There is evidence that an allele at a major locus elevates the rate of sodium-lithium counter transport (Hasstedt 1988). One study presented evidence of genetic linkage between the angiotensinogen gene (AGT) and hypertension in humans (Jeunemaitre 1992). The study demonstrated association of AGT molecular variants with the disease and found significant differences in the plasma concentrations of angiotensinogen among hypertensive subjects with different AGT genotypes.

The hypertension group includes hypertensive heart and renal disease. The GIF score for all the cases was 3.04 (P = .0129), and for the younger (under age 73) set it was 3.26 (P = .0467). The relative risk for all the deaths was 1.45 (P = 0), and for the younger set it was 2.18 (P = 0). The younger cases show a familial aggregation for hypertension related deaths.

The **heart disease** group includes ischemic heart disease, angina pectoris, and coronary atherosclerosis. Since heart disease is closely related to **myocardial infarctions** (destruction of heart tissue cause by inadequate supply of oxygen), they will both be discussed here. Coronary heart disease is the leading cause of death in the United States and most of the industrialized western world. The risk factors include hypertension, cigarette smoking, elevated serum cholesterol, genetic susceptibility, and gender. There is abundant evidence that cigarette smoking and obesity increase the risk of heart disease, whereas exercise and a diet low in saturated fat decreases the risk. Mortality from heart disease is higher in men than in women under the age of 50 and tends to equal out after the age of 50.

Many studies have examined the role of family history in heart disease and they show that an individual is two to seven times more likely to suffer from heart disease than an individual with no family history. Generally these studies show that the risk increases if there are more affected relatives, the affected relative is female or the age of onset of the affected relative is less than 55 (Jorde 1995).

A number of genes contribute to heart disease such as the eight apolioprotein genes and the LDL receptor gene. The apolipoprotein gene Apo A-I is linked to familial hypoalphalipoproteinemia, which is the most common form of primary depression of HDL-cholesterol (Third 1984). HDL is thought to be beneficial in preventing coronary heart disease. Defects in the LDL receptor gene are believed to be responsible for familial hypercholesterolemia (FH), which is characterized by the elevation of serum cholesterol bound to low-density lipoprotein (LDL). Elevated levels of LDL are a risk factor for heart disease. Lipoprotein measurements may help predict the risk of coronary heart disease in individuals with FH (Houlston 1988). FH is an important cause of heart disease, accounting for approximately 5% of myocardial infarctions in persons under 60. FH is one of the most common autosomal dominant disorders with about 1 in 500 persons a heterozygote.

The only significant GIF for heart disease was in the younger (under age

71) set where it was 3.08 (P = .0074) The GIF for all deaths from myocardial infarction was 3.07 (P = .0001), for the younger (under age 69) cases it was 3.44 (P = 0), and for males it was 3.09 (P = 0). There were almost twice as many male deaths as female deaths. The GIF for both causes of death combined was 3.14 (P = 0). The relative risk for all deaths from heart disease was 1.22 (P = 0), and for the younger cases it was 1.51 (P = 0). For myocardial infarction, the relative risk of all the deaths was 1.46 (P = 0), and for the younger cases it was 1.80 (P = 0). There was evidence of familiality for these two causes of death, especially when the age of death is younger. The average age of death for both myocardial infarction and coronary heart disease was more than 73. It is likely that there are many sporadic cases that are a result of advanced age.

Endocarditis ensues when bacteria entering the blood stream from an oral or other source lodge on heart valves that may already bear platelet-fibrin thrombi. The frequency of bacteremia is quite high after dental extraction or periodontal surgery. Rheumatic heart disease and congenital heart disease are predisposing factors. The infection may cause rupture of the valve tissue itself or of is chordal structures, leading to either gradual or acute valvular regurgitation. Some other effects of the infection include the formation of emboli in the heart and the disruption of the heart conduction.

The GIF of all the cases of endocarditis was 3.17 (P = .2365), which was not significantly different from the controls. The highest GIF score was 3.90 (P = .1120) from the males. This may be an indication of a male susceptibility to congenital heart disease, which does show familial tendencies. There were not enough first-degree relatives with endocarditis to calculate a relative risk.

There are three classifications of **cardiomyopathy**: dilated, hypertrophic, and restrictive. In dilated cardiomyopathy, ventricular enlargement occurs and systolic dysfunction results in symptoms of congestive heart failure. The cause of dilated cardiomyopathy appears to be the end result of myocardial damage produced by a variety of toxic, metabolic, and infectious agents. A number of studies have shown familial aggregation of dilated cardiomyopathy and a possible autosomal dominant inheritance pattern (Gardner 1987; MacLennan 1987).

Hypertrophic cardiomyopathy is characterized by myocardial hypertrophy, especially involving the interventricular septum. Dyspnea is the most common symptom. In many patients, the disease appears to be transmitted genetically as an autosomal dominant disorder with a high degree of penetrance, but sporadic cases do occur. This pattern of inheritance was confirmed in a study of 50 families (Greaves 1987).

Restrictive cardiomyopathies are less common than the other two types. They are caused by a variety of infiltrative processes, including amyloidosis, hemochromatosis, sarcoidosis, endomyocardial fibrosis, and endocarditis. Restrictive cardiomyopathy is characterized by abnormal diastolic function that impedes ventricular filling. A possible genetic link is hemochromatosis which has been previously discussed.

The GIF of all deaths from cardiomyopathy was 3.75 (P = .0204). It was substantially higher in the younger (under age 67) cases where it was 5.85 (P = .0152). The relative risk of all cardiomyopathy deaths was 2.74 (P = .1056). There was a small sample set, so there were not enough first-degree relative cases to calculate a relative risk score for the younger set. The results do show a definite familial predisposition for cardiomyopathy.

Heart conduction disorders are a result of a number of diseases involving the heart, such as coronary heart disease, endocarditis, and cardiomyopathy. The genesis of cardiac arrhythmias is divided into disorders of impulse formation, impulse conduction, and combinations of the two. Conduction disorders can lead to dizziness, palpitations, congestive heart failure, and sudden death. More severe outcomes are common in patients with diseased hearts.

Both the GIF results and the relative risk scores do not show any familial tendencies. This is most likely due to the number of factors that can cause conduction disorders.

Heart failure refers to a state in which the heart cannot provide sufficient cardiac output to satisfy the metabolic needs of the body. It is commonly called **congestive heart failure**, as symptoms of increased venous pressure (pulmonary congestion with left heart failure and peripheral edema with right heart failure) are often prominent.

Congestive heart failure can result from several diseases. The most common in western industrialized countries are atherosclerotic coronary artery disease and myocardial infarction. Myocarditis, cardiomyopathy, and valvular and congenital defects can result in heart failure. Mitral and aortic regurgitation and ventricular and atrial septal defects cause volume overload states; aortic and pulmonic stenosis and hypertension cause pressure overload states. Conditions that restrict ventricular filling, such as mitral stenosis, constrictive pericarditis, or restrictive cardiomyopathies, cause heart failure.

Since many of the causes of congestive heart failure (CHF) have familial tendencies, one would expect CHF to show familial aggregation. The GIF for all deaths from CHF was 3.28 (P = .0173). In the younger (under age 79) cases it was 3.91 (P = .0011). The relative risk for CHF was 1.35 (P = .1685), and it was 1.90 (P = .0885) in the younger cases. There is a familial tendency in all the cases that is stronger in the younger cases.

Cerebrovascular diseases include disorders of the arterial or venous circulatory systems that produce or threaten to produce injury to the central nervous system. The general term **stroke** describes the functional neurologic injury. Stroke takes a worldwide toll, especially affecting persons over the age of 55. Although the incidence has declined in recent years, only heart disease and cancer exceed stroke as causes of death and disability in developed countries. A number of risk factors for strokes are both environmental and familial. The major risk factors for stroke are hypertension, smoking, atrial fibrillation, myocardial infarction, hyperlipidemia, diabetes, congestive heart failure, and acute alcohol abuse.

Several studies that focused on hereditary multi-infarct dementia in multiple members of families found a pattern consistent with autosomal dominant inheritance (Sourander 1977; Sonninen 1987). The GIF results and the relative risk scores show a slight familial influence in stroke. The fact that there is not a larger familial component is most likely due to the number of different risk factors. The GIF for all the cases of stroke was 3.16 (P = .0003). It was 3.41 (P = 0) in males, which was the highest GIF for stroke. The relative risk was 1.30 (P = 0) for all cases, and 1.39 (P = .0002) for males.

Aortic **aneurysms**, localized areas of increased diameter of the aorta, may occur in the ascending aorta, aortic arch, descending thoracic aorta, or abdominal aorta, depending on the etiology. Risk factors of aortic aneurysms include Marfan's syndrome, syphilis, endocarditis, congenital lesion, and atherosclerosis.

Intracranial aneurysms occur in three forms: fusiform, mycotic, and congenital "berry" aneurysms. Fusiform aneurysms represent ectatic dilatations of the basilar or intracranial portion of the carotid artery. Usually they produce no symptoms, but sometimes their large size compresses adjacent tissues or cranial nerves to cause local neurologic dysfunction. Mycotic aneurysms arise in the course of bacterial endocarditis when septic emboli lodge in a peripherally located cerebral vessel. Congenital berry aneurysms arise at the base of the brain. Berry aneurysms are thought to result from a congenital defect that affects adventitial tissue and muscle at arterial branch points along the base of the brain. Congenital aneurysms are more common in individuals with long-standing hypertension. Most intracranial aneurysms are detected only when they rupture, an event that can occur at any age but most commonly occurs between the age of 40 and 65.

A number of studies discussing familial aneurysms have been done. One study reported a 10% incidence of familial intracranial aneurysms (Ronkainen 1993). Possible defects in type III collagen was mentioned as a cause of familial multiple intracranial aneurysms (De Paepe 1988). A review of the literature of familial intracranial aneurysms found 238 families with 560 affected members, of which 56% were female and 44% were male (Schievink 1994). The most common affected kinship was among siblings.

Aneurysm showed one of the strongest familial tendencies of all of the causes of death. The GIF for all cases was 4.02 (P = 0), and for the younger (under age 70) cases it was 6.85 (P = 0). The GIF for females at 6.48 (P = .0001) was almost two times as large as the male GIF. The relative risk for all cases was 3.87 (P = 0), and for the younger cases it was 6.70 (P = 0). The relative risk was also substantially higher for females than males. Aneurysms appear to be strongly familial.

There are a number of different types of **circulatory disorders** such as inflamed arteries and blood clots. There were a small number of deaths from these diseases. The GIF for all cases did not show a significant familial influence, but the GIF for male deaths, which was 4.90 (P = .0303), did show familial influence. One study described a large kindred with cytopenia and occlusive vascular disease. Vascular occlusive disease occurred in 9 of 13 adults. Both males and females were affected and male-to-male transmission was observed (Aufderheide 1972).

Pulmonary embolism is most commonly caused by the embolic material thromboemboli. Consequences of the thromboembolis depend on the amount of clot reaching the lung and the pulmonary condition of the patient. The consequences may vary from a persistent tachycardia or mild dyspnea to cardiopulmonary arrest. Thromboemboli directly or indirectly cause 200,000 deaths per year. Medical risk factors include cancer, stroke, myocardial infarction, congestive heart failure, pregnancy, and sepsis. Other risk factors include orthopedic surgery, lower extremity fractures, and major surgery.

Inherited risk factors include protein C deficiency, antithrombin III and plasminogen activation disorders. Protein C is a vitamin K-dependent serine protease zymogen, and it has an important anticoagulant role. The deficiency of protein C and its relationship to thromboembolic diseases were first found in a kindred where affected members had low levels of the plasma protein C antigen (Griffin 1981). Clinically unaffected members of the kindred had normal levels. In a large New England kindred, a strong statistical correlation was found between thromboembolic disease and protein C deficiency (Bovill 1989). Antithrombin deficiency in individual patients with severe venoocclusive disease, along with a complete family history, was also reported (Nesje 1970).

The GIF and the relative risk for pulmonary embolism show a definite familial aggregation. The GIF for all the cases was 3.57 (P = .0125), and for the younger (under age 70) cases it was 4.38 (P = .0086). The relative risk for all the cases was 2.33 (P = .0392), and for the younger cases it was 3.41 (P = .0427). The relative risk of 3.04 (P = .0352) for males was double the risk of females.

Pneumonia currently accounts for about 10% of admissions to adult medical services in North America and is the sixth leading cause of death in the United States. A number of pathogens cause pneumonia such as streptococcus pneumoniae, mycoplasm pneumoniae, influenza virus, and mycobacterium tuberculosis. Certain systemic disorders are associated with pneumonia due to particular organisms. These include seizures, alcoholism, diabetes, sickle cell disease, chronic lung disease, and chronic renal failure.

Familial aggregation could come from shared environment and exposure to the same pathogens. Some studies have shown there is a genetic factor in the actions of immune response antigens (Hsu 1981; Meyer 1994). There is also significant familial tendencies in some cases of the disease when it is associated with diseases such as diabetes and alcoholism.

The GIF of all the deaths from pneumonia was only 2.97 (P = .1062), but the GIF for the younger (under age 75) cases was 3.27 (P = .0073). For men the GIF was 3.08 (P = .0229). There was no significant relative risk, as the relative risk for all the pneumonia groups was close to one. The GIF values show a slight familial influence.

There were a small number of **influenza** deaths. The influenza virus is the most common virus that causes pneumonia. Viral pneumonia typically occur in community epidemics. The influenza virus weakens individuals, so that they become prone to other infections. Yearly immunization with the influenza vaccine decreases morbidity and mortality due to secondary bacterial pneumonia. The family influence of shared environment would be a likely cause of influenza, but there is research mentioned with pneumonia that introduced evidence for genetic influence on immune responses.

The GIF for influenza was high with the value for all cases 4.37 (P = .0563) and the value for males 7.57 (P = .0131). There were not enough cases for the relative risk calculation. The high GIF values could be attributed to inherited deficiencies of the immune system. The contribution to GIF from extended family for influenza is evident when compared to the control sets. This supports the belief that inherited aspects of the immune system do have an effect with a common disease like influenza.

Bronchitis is associated with emphysema, bronchospasm, and airway obstruction. As with emphysema, cigarette smoke is the major risk factor, although exposure to other airborne pollutants may play a role by causing chronic irritation. Both the GIF and the relative risk calculations did not show any familial aggregation for bronchitis.

Asthma is characterized by airway obstructions that vary over time and is completely or partially reversible with treatment. Acute severe asthma refers to an attack of increased severity which is unresponsive to routine therapy and which can lead to death. The airways are the site of an inflammatory response consisting of cellular infiltration, epithelial disruption, mucosal edema, and mucosal plugging. The stimulus for the inflammation may be immunologic in origin, as is the case in classic extrinsic asthma, in which mast cells, sensitized by IgE antibodies, degranulate and release bronchoactive mediators following exposure to a specific antigen.

Possible inherited factors for asthma could involve the immune system. Also an arachidonate metabolite is a constrictor of vascular and smooth respiratory muscles (Ushikubi 1989). It has been implicated as a mediator in bronchial asthma.

The GIF for all deaths from asthma was 3.97 (P = .0640), and the GIF for the younger (under age 70) cases was 7.57 (P = .0004). These show a strong familial influence for asthma. There were not enough first-degree relative deaths from asthma to calculate the relative risk.

Patients with **chronic airway obstruction** have slowly progressive airway obstruction. The course of the disease is punctuated by periodic exacerbations resulting in an increase in dyspnea and sputum production or, occasionally, the precipitation of acute respiratory failure. Chronic airway obstructions generally affect middle-aged and elderly individuals. Three pathophysiologic disorders are associated with chronic airway obstruction: emphysema, small airways obstruction, and chronic bronchitis. The familial tendencies of emphysema are discussed below.

The GIF values and the relative risk scores for chronic airway obstruction show a definite familial aggregation, especially in women. The GIF for all cases was 3.95 (P = 0), and for the younger (under age 70) cases it was 5.22 (P = .0004). The GIF was also high for females at 5.01 (P = .0073). The relative risk for all cases was 2.98 (P = 0), for the younger cases it was 3.82 (P = .0010), and for the females it was 4.63 (P = .0001).

Emphysema is characterized by two features. Anatomically, it is defined as an abnormal enlargement of the air spaces distal to the terminal bronchiole, accompanied by destructive changes in the alveolar walls. Physiologically, it is characterized by a loss of elastic recoil and thus an increase in lung compliance.

Most researchers believe that emphysema is caused by an imbalance of protease and antiprotease in the lung, with the resultant lung destruction. This theory is based on the discovery of a small number of patients with an inherited deficiency of alpha-antiprotease, the major alphaprotease, which develops without any other risk factors. Cigarette smoke, the major risk factor for emphysema, has been shown to increase the number of alveolar macrophages and neutrophils in the lung. It also enhances protease release and impairs the activity of antiprotease. However, other factors must determine susceptibility to emphysema, because fewer than 10 to 15% of smokers develop clinical evidence of airway obstruction. Familial emphysema has been reported in a number of studies (Larsen 1965; Hole 1965; Knudsen 1979).

Since the smoking rate in Utah is relatively low, the GIF and relative risk values should be free of some of the environmental risk factors of emphysema. The GIF for all the emphysema deaths was 3.72 (P = .0019), and for the younger (under age 65) cases it was 8.00 (P = 0). There were almost 10 times as many male deaths as female deaths. The relative risk for all the cases was 1.70 (P = .0228), and for the younger cases it was 3.07 (P = .0062). These values show a significant familial tendency for emphysema. The numbers are similar to those from chronic airway obstruction, which is related to emphysema.

The **ulcer** group included gastric, duodenal, and peptic ulcers. The lifetime prevalence of peptic ulcer disease is 5 to 10%, with about equal prevalence in men and women. Duodenal ulcers are more frequent than gastric ulcers. The incidence of ulcer disease increases with age. Genetic factors seem to be important in some patients with peptic ulcers. There is an increased incidence of duodenal ulcer in families that is related to the autosomal dominant transmission of elevated serum pepsinogen (Rotter 1982). Other risk factors include smoking, ethnic background, the use of nonsteroidal anti-inflammatory drugs and various diseases such as chronic lung disease, cirrhosis, and chronic renal failure.

The GIF for all the cases of ulcer deaths was 3.21 (P = .2156), and for the younger (under age 70) cases it was 4.03 (P = .1322). It was also high in females at 3.92 (P = .1086). The relative risk for all the cases was 1.56 (P = .2119), and for the younger cases it was 2.71 (P = .1075). For the women it was 2.83 (P = 0). There seems to be some familial aggregation for ulcers that is strongest in females.

Intestinal obstruction can be caused by mechanical obstructions such as ulcers or tumors. Other causes are certain drugs, electrolyte imbalance, metabolic disorders, neuromuscular disorders, brain stem tumors, and psychiatric disorders. The GIF values showed no familial predisposition.

Colonic **diverticulosis** is characterized by saccules of mucosa covered by serosa. It develops commonly in later life, particularly in western societies. The formation of diverticula is believed to be caused by any condition that chronically increases intraluminal pressures, such as a low-fiber diet. They become clinically important if they bleed.

Since diet is such an important risk factor, any familial aggregation is likely environmental. The GIF for all the cases is 3.73 (P = .2665), which was not a significant difference from the controls. The GIF for younger (under age 75) cases was 6.20 (P = .1710). There seems to be a familial predisposition, but there was a small number of deaths from diverticulosis, so the p-values for the GIF results are high.

Cirrhosis is the irreversible end result of fibrous scarring and hepatocellular regeneration that constitute the major responses of the liver to a variety of long-standing inflammatory, toxic, metabolic, and congestive insults. Some of the major complications of cirrhosis are portal hypertension, liver failure, and hepatocellular carcinoma. Alcohol abuse and hepatitis C are the most common causes of cirrhosis in the western world, whereas hepatitis B is a major cause in the third world. Other causes of cirrhosis include hemochromatosis and

Wilson's disease. Hemochromatosis is a genetically determined iron storage disorder that was discussed previously with liver cancer. Wilson's disease is a rare autosomal recessive disorder characterized by a defect in hepatic excretion of copper.

Since most of the causes of cirrhosis run in families, either for genetic reasons or for environmental reasons, a familial predisposition for cirrhosis is expected. The GIF for all cases of cirrhosis was 3.33 (P = .1042). A similar value was calculated with the younger cases, but a higher GIF of 3.85 (P = .0583) was calculated with the females. A similar pattern was seen in the relative risk scores where the relative risk for all the cases was 2.01 (P = .0594) and the relative risk for females was 2.58 (P = .0655). Cirrhosis does show a familial tendency.

Biliary tract and gallbladder disorders have some of the same outcomes as cirrhosis. Biliary disorders often lead to cirrhosis in women. The most serious gallbladder disorder is acute cholecystitis which is caused by obstruction of the cystic duct. It leads to distension, inflammation, and secondary infection of the gallbladder. The mortality of acute cholecystitis is 5 to 10% and is almost entirely confined to patients over 60 years of age. Neither the GIF nor the relative risk results showed a significant familial influence for these disorders.

Nephritis has a number of different forms. It is characterized by the onset of hematuria and proteinuria temporally associated with the development of hypertension. Glomerulonephritis is caused by bacterial infections, endocarditis, and viral infections. It affects the glomerulus, which is a capillary bed through which blood flows in and out of the kidneys. Hypertension and edema are features of glomerulonephritis.

Tubulointerstitial nephropathy encompasses a group of clinical disorders that affect the renal tubules and interstitium principally, with relative sparing of the glomeruli and renal vasculature. Acute interstitial nephritis is caused by complications of a wide variety of drugs, especially antibiotics and nonsteroidal anti-inflammatory drugs. The major clinical manifestation of acute interstitial nephritis is the development of acute renal insufficiency.

Hereditary nephritis (Alport's syndrome) usually presents in childhood with recurrent gross hematuria. Sensorineural deafness is present in about 50% of the patients. Family history may reveal any number of different patterns, although most pedigrees show some X linkage. Males are usually affected more than females and often develop renal failure before age 30. As many as six different types of Alport's syndrome were identified in families based on mode of inheritance, age of onset, and severity (Atkin 1986). The identification of mutations in the COL4A5 collagen gene was linked to Alport's syndrome (Barker 1990).

The GIF and relative risk results show a familial aggregation for nephritis. The GIF for all the cases was 3.56 (P = .0785). For the younger (under age 65) cases it was 4.23 (P = .1082), and for the males it was 4.10 (P = .0493). The relative risk for all the cases was 3.23 (P = .0301), and for the younger cases it was 5.87 (P = .0314).

Renal failure is often caused by nephritis. Other causes include hypertensive nephrosclerosis, diabetes, and cystic kidney disease. Kidney failure is associated with the accumulations of potentially toxic substances in the body. Some of the most common disorders associated with renal failure are cardiovascular disease, anemia, endocrine dysfunction, and neurologic complications. Treatment options include dialysis and transplantation.

The GIF and the relative risk results did not show any familial aggregation for renal failure. This is likely due to the wide variety of disorders that lead to renal failure.

Congenital anomalies of the circulatory system consist mainly of congenital heart disease which refers to cardiac lesions present at birth. Congenital heart disease results from both genetic and environmental factors. Congenital heart disease may be familial in some instances, but a distinct pattern has not been recognized. It is more common in children of older mothers and premature infants. Environmental factors such as teratogens and maternal rubella are commonly recognized risk factors.

The GIF and relative risk results show a familial aggregation of congenital heart disease. The GIF for all the cases was 5.10 (P = 0). It was especially high in males at 7.47 (P = 0), whereas in females it was less than the controls. The relative risk for all the cases was 7.60 (P = .0359).

Suicide is the second leading cause of death among young men in Utah. There were more than four times as many male deaths from suicide as there were female deaths. One study has linked a variant human brain specific protein to depression and suicide (Comings 1979). Researchers have shown that two other psychiatric disorders, schizophrenia and manic-depressive disorder, aggregate in families.

The GIF and relative risk results show that suicide has a strong familial aggregation in the linked death certificate records. The contribution to the GIF graph that was plotted for suicide does show that most of the familial influence comes from close relatives and not extended family members. The GIF for all deaths by suicide was 3.71 (P = 0). For the younger (under age 40) cases it was 4.75 (P = 0), and for males it was 4.02 (P = 0). The female GIF was less than the controls. The relative risk for all the cases was 4.64 (P = 0), and for the younger cases it was 4.80 (P = 0). The male relative risk was 4.76 (P = 0).

Motor vehicle accident deaths were studied as a cause of death that would not have any genetic predisposition. The results did show some familial aggregation, which is likely due to multiple family members dying in the same accident. The GIF for motor vehicle accidents was 3.11 (P = 0). For the younger (under age 40) cases it was 3.6 2(P = 0), and for the males it was 3.23 (P = .0001). The relative risk for all the deaths was 2.09 (P = 0), and for the younger cases it was 2.85 (P = 0). There were twice as many male deaths as female deaths. The average age of death was also one of the lowest at 48.1 years.

4.9 Conclusions

The linked death certificates proved to be a valuable source of information for the study of familiality in common diseases. The results of the genetic epidemiological analysis supported much of the current research in the genetics of common diseases. Information was found for causes of death that have not been studied extensively. Confidence in the results was highest for the most common causes of death. Sample size was important in establishing the reliability of the GIF and relative risk.

Another factor that could influence the reliability of the results is the correct determination of cause of death for each individual. There are possibly cases where the cause of death was the result of a disease that was not listed on the death certificate. For example, complications from diabetes often lead to a number of medical disorders that could be fatal. There would be no way of knowing if the individual had diabetes, unless it was listed as a secondary cause of death. Only the more recent death certificates had secondary causes of death coded. This could also affect cancer diagnosis where a primary tumor metastasizes to another site. The secondary site could be listed as the cause of death without mention of the primary site.

The two methods to analyze the linked death certificates use different approaches to look for familiality. The Genealogical Index of Familiality looks at all of the possible relationships for cases with the same cause of death. It examines both the close relationships and the extended relationships. The examination of distant relationships helps to distinguish between environmental and genetic causes for common diseases. The first-degree relative risk determines the familiality by comparing the rates of the disease in first-degree relative to the rate of a complete population, which was all of the individuals in the genealogy who had died in Utah. This is in contrast to the GIF that looks at the incidence of kinship between the death certificates and controls selected from the genealogy records. The two different approaches aid in determining which causes of death can be attributed to genetic predispositions.

For some causes of death, considerable genetic evidence has been identified. This is true for breast cancer, colon cancer, ovarian cancer, melanoma, heart disease, Alzheimer's disease, and hypertension. The death certificates produced results that clearly showed the familiality for most of these diseases, especially in the sets of the youngest cases. One of the exceptions was melanoma which had a small sample size. There are more than a 1000 individuals with melanoma in the Utah Cancer Registry, who are linked to a genealogical record whereas there were less than 400 death certificates with melanoma linked to the genealogy. Perhaps the analysis of the deaths from melanoma did not show any familiality because the mortality from the disease is far lower than the frequency of the disease.

Familiality that can be attributed to genetic predisposition is evident for leukemia, diabetes, obesity, aneurysms, cardiomyopathy, congestive heart failure, congenital anomalies of the circulatory system, and multiple sclerosis. Hereditary defects in the immune system make individuals susceptible to a number of diseases. This is possibly a factor in the familiality seen in myeloma, lymphoma, leukemia, and influenza. There are a number of diseases where it is speculated that a genetic susceptibility to an environmental agent trigger the disease. This could be a genetic reason for the high familiality observed in kidney cancer, stomach cancer, bladder cancer, and Parkinson's disease.

A common genetic predisposition for emphysema, chronic airway obstruction, and asthma could exist since the three diseases affect the same physiologic system. All three diseases showed a high degree of familiality. Other causes of death that showed high familiality that could be genetically related include alcohol-related deaths, motor neuron diseases, congestive heart failure, pulmonary embolism, and Hodgkin's disease.

There were some diseases where there was evidence of familiality, but there is some question of whether the cause is environmental or genetic. These include lung cancer, suicide, ulcers and diverticulosis. Motor vehicle accidents were also familial but are obviously not genetic. Pneumonia and pancreatic cancer were two of the most common causes of death that did not show any familiality.

The ability to combine the death certificates and the genealogy records from the UPDB has provided an interesting examination of familiality in common causes of death. The genetic analysis of the record links has confirmed much of the current knowledge of the genetic predispositions to common diseases. It has also introduced new areas where further research would be warranted. The familiality research has shown the value of record linking.

CHAPTER 5

SUMMARY AND DISCUSSION

5.1 Record Linking

The probabilistic approach to linking the genealogy, cancer, and death certificate records worked well. It was able to make use of all the information available for linking. The Automatch software was an adequate tool to use with the data sources, despite data from the cancer registry, the death certificates, and the genealogy database that were inconsistent and often incomplete. It would provide a standard approach to a variety of record-linking applications.

A number of Automatch tools were of great use such as the histograms created during each match run that helped to choose cutoff values for matches. Another useful tool was the **mprob** program that calculated the *m* probabilities used in the calculations of the linking weight for each field.

A problem in the matching step is the method used for the calculation of the *u* probability or the frequency of a variable. Automatch assigns a unique weight for only the 100 least frequent values of a variable. This is appropriate for a field such as age or birth year where there are not many unique values, but for a field such as last name, there are certainly more than 100 names that are unique and should be scored higher than common names such as Smith or Jones. This causes the program to lose some of the discriminating power of a name.

One piece of the software that could be improved was the report program. It would only produce reports for all of the matching passes. It was difficult to analyze each matching pass when a large report covering all the matching passes was produced. It would make the selection of matching cutoffs easier if reports for each matching pass could be created.

The record linking results showed the value of complete information for each record. The maiden name field for females contributed greatly to finding a match when it was part of the record. There was also a substantial difference in percentage of records linked for the years in which the death certificates lacked a birth year. This shows how much the loss of information from one field of the record can affect the linking outcome.

5.2 Cause of Death Study

The record linking produced a large set of linked records for the cause of death study. The 126,085 linked records made it possible to do a comprehensive study of the familiality for a large number of common causes of death. Since the same methods of analysis were applied to each cause of death, it was possible to rank the familiality of each cause of death and show which causes of death were the most familial.

The first-degree relative risk and genealogical index of familiality were useful methods for determining the familiality of the causes of death. The GIF does give a better indication of genetic predispositions since it looks at extended relatives whereas the relative risk only looked at first-degree relatives. Other informative relative risk studies could look at the risk of second-degree relatives and the spouses of individuals from the death certificates. Spouses would share the same environments, but since they would have a different genotype the environmental effect could be compared to possible genetic predispositions.

There are problems in determining genetic predispositions from the death certificates. There are a number of diseases such as diabetes, melanoma and colon cancer where there is substantial evidence of genetic predisposition. The degree of familiality shown in the death certificates was not very strong for these diseases. Some reasons for this could be age of onset of the disease or the disease leading to other causes of death. Another factor that could influence that low familiality values for colon cancer and melanoma is the increased awareness of the disease in a family. A death in a family from a cancer that is genetically predisposed may help to prevent future deaths. Preventative measures such as routine screening and diet changes have shown to reduce the number of deaths from certain diseases.

Such interventions show the value of finding genes which predispose common diseases and developing diagnostic tests for the genes. The death certificates showed a strong degree of genetic predisposition for ovarian cancer. Ovarian cancer is often found when it is difficult to treat. If a woman were to know that she had a strong chance of developing ovarian cancer, she could take preventative measures that could reduce the chances of developing the disease.

When the analysis was stratified by age of death, the strongest evidence of familiality was often seen in the youngest third of each cause of death. These data sets are the most likely indicator of a genetic predisposition for most of the causes of death and would warrant further study. Younger age of death limits could be tried for those causes of death where there was a large sample size such as myocardial infarction, heart disease, and diabetes. It is also interesting to note causes of death such as alcohol-related deaths or suicide, where the number of males greatly outnumber the females. There seem to be either genetic or environmental causes that only affect males that could also be studied further.

The death certificate analysis produced results that were consistent with much of the current research in the study of genetics for common diseases. The death certificate analysis also found strong familial aggregation in a number of diseases where little is known about genetic predispositions and where further study could be done. These include kidney cancer, stomach cancer, chronic airway obstruction, aneurysm, and emphysema.

Record linking has a place in medical and genetic research and will become a valuable tool as more data repositories and registries are created. The cause of death studies produced many interesting results that showed the value of the record linking project. A valuable resource for genetic, demographic and epidemiological studies was created with the large set of linked death certificate records.

REFERENCES

Anderson, D. E., J. L. Smith, and C. M. McBride. 1967. Hereditary aspects of malignant melanoma. Journal of American Medical Association 200: 741-746.

Anderson, K. C., D. S. Jamison, W. P. Peters, and F. P. Li. 1986. Familial Burkitt's lymphoma: association with altered lymphocyte subsets in family members. American Journal of Medicine 81: 158-162.

Andreoli, T. E., J. C. Bennett, C. J. Carpenter, F. Plum, and L. H. Smith, eds. 1993. Essentials of medicine. Philadelphia: W.B. Saunders Company.

Arason, A., R. B. Barkardottir, and V. Egilsson. 1994. Linkage analysis of chromosome 17q markers and breast-ovarian cancer in Icelandic families, and possible relationship to prostatic cancer. American Journal of Human Genetics 52: 711-717.

Atkin, C. L., M. C. Gregory, and W. A. Border. 1986. Alport syndrome. in Strauss and Welt's diseases of the kidney fourth edition, eds R. W. Schrier and C. W. Gottschalk. Boston: Little, Brown.

Auferheide, A. C. 1972. Familial cytopenia and vascular disease: a newly recognized autosomal dominant condition. Birth Defects Original Article Service 8: 63-68.

Baldwin, J. A., E. D. Acheson, and W. J. Graham, eds. 1987. Textbook of medical record linkage. Oxford: Oxford University Press.

Barbeau, A., T. Cloutier, M. Roy, L. Plasse, S. Paris, and J. Poirier. 1985. Ecogenetics of Parkinson's disease: 4-hydroxylation of debrisoquine. Lancet II: 1213-1216.

Barker, D. F., S. L. Hostikka, J. Zhou, L. T. Chow, A. R. Oliphant, S. C. Gerken, M. C. Gregory, M. H. Skolnick, C. L. Atkin, and K. Tryggvason. 1990. Identification of mutations in the COL4A5 collagen gene in Alport syndrome. Science 248: 1224-1227.

Bishop, D. T. and M. H. Skolnick. 1984. Genetic epidemiology of cancer in Utah genealogies: a prelude to the molecular genetics of common cancers. Cell Physiology, Suppl: 63-77.

Bovill, E. G., K. A. Bauer, J. D. Dickerman, P. Callas, and B. West. 1989. The clinical spectrum of heterozygous protein C deficiency in a large New England kindred. Blood 73: 712-717.

Buehler, J. W., R. L. Berkelman, and J. W. Curran. 1989. Reporting of AIDS: tracking HIV morbidity and mortality. Journal of American Medical Association 262: 2896-2897.

Cannon, L., D. T. Bishop, M. Skolnick, S. Hunt, J. L. Lyon, and C. R. Smart. 1982. Genetic epidemiology of prostate cancer in the Utah mormon genealogy. Cancer Survey 1: 48-69.

Cannon-Albright, L. A., D. T. Bishop, C. Goldgar, and M. H. Skolnick. 1991. Genetic predisposition to cancer. In Important advances in onclogy 1991. eds V. T. Devita, S. Hellman, and S. A. Rosenberg, 39-55. Philadelphia: J. B. Lippencott.

Cannon-Albright, L. A., D. E. Goldgar, L. J. Meyer, C. M. Lewis, D. E. Anderson, J. W. Fountain, and M. E. Hegi, et al. 1992. Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. Science 258: 1148-1152.

Cannon-Albright, L. A., M. H. Skolnick, T. Bishop, R. G. Lee, and R. W. Burt. 1988. Common inheritance of susceptibility to colonic adenomatous polyps and associated colorectal cancers. New England Journal of Medicine 319 (September): 533-537.

Cannon-Albright, L. A., A. Thomas, D. E. Goldgar, K. Gholami, K. Rowe, M. Jacobsen, W. P. McWhorter, and M. H. Skolnick. 1994. Familiality of cancer in Utah. Cancer Research 54 (May): 2378-2385.

Carter, B. S., T. H. Beaty, G. D. Steinberg, B. Childs, and P. C. Walsh. 1992. Mendelian inheritance of prostate cancer. Procedings National Academy of Science 89: 3367-3371.

Casey, G., Y. Yamanaka, H. Freiss, M. S. Korbrin, M. E. Lopez, M. Buchler, H. G. Beger, and M. Korc. 1993. p53 Mutations are common in pancreatic cancer and are absent in chronic pancreatis. Cancer Letters 69: 151-160.

Chissoe, S. L., A. Bodenteich, Y. F. Wang, Y. P. Wang, D. Burian, S. W. Clifton, J. Crabtree, and A. Freeman. 1995. Sequence and analysis of the human ABL gene, the BCR gene, and the regions involved in the Philadelphia chromosomal translocation. Genomics 27: 67-82.

Cloninger, C. R. 1987. Neurogenetic adaptive mechanisms in alcoholism. Science 236: 410-416.

Cohen, A. J., F. P. Li, S. Berg, D. J. Marchetto, S. Tsai, S. C. Jacobs, and R. Brown. 1979. Hereditary renal-cell carcinoma associated with chromosomal translocation. New England Journal of Medicine 301: 592-595.

Cohen, M.M., and P. G. Duncan. 1988. Physical status score and trends in anesthetic complications. Journal of Clinical Epidemiology 41: 83-90.

Comings, D. E. 1979. Pc1 Duarte, a common polymorphism of a human brain protein and its relationship to depressive illness and multiple sclerosis. Nature 277: 28-32.

Comotti, B., R. Bassan, M. Buzzetti, G. Finazzi, and T. Barbui. 1987. Multiple myeloma in a pair of twins. British Journal of Haematology 65: 123-124.

Conley, C. L., J. Misiti, and A. J. Laster. 1980. Genetic factors predisposing to chronic lymphocytic leukemia and to autoimmune disease. Medicine 5: 323-334.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximun likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39: 1-38.

De Paepe, A., W. Van Landeghem, F. De Keyser, F. M. Pope, and M. Matton. 1988. Collagen type III deficiency associated with multiple intracranial aneurysms. Clinical Genetics 33: 462.

Donald, J. A., W. Barendse, and D. W. Cooper. 1989. Linkage studies of HLA and insulin gene restriction fragment length polymorphisms in families with IDDM. Genetic Epidemiology 6: 77-81.

Durst, M., S. C. Amsbaugh, and J. A. Dipaolo. 1987. Human papillomavirus type 18 DNA is integrated at a single chromosome site in cervical carcinoma cell line SW756. Journal of Virology 51: 1682-1685.

Ebers, G. C., A. D. Sadovnick, and N. J. Risch. 1995. A genetic basis for familial aggregation in multiple sclerosis. Nature 377: 150-151.

Edwards, C. Q., G. E. Cartwright, M. H. Skolnick, and D. B. Amos. 1980. Genetic mapping of the hemochromatosis locus on chromsome six. Human Immunology 1: 19-22.

Farrer, L. A., R. H. Meyers, L. A. Cupples, P. H. St. George-Hyslop, T. D. Bird, M. N. Rossor, M. J. Mullan, R. Polinsky, and L. Nee. 1990. Transmission and age-at-onset patterns in familial Alzheimer's disease: evidence for heterogeneity. Neurology 40: 395-403.

Fellegi, I. P., and A. B. Sunter. 1969. A theory for record linkage. Journal of the American Statistical Association 64: 1183-1210.

Francis, D. A., J. R. Batchelor, W. I. McDonald, I. A. Dodi, S. N. Hing, J. E. Hern, and A. W. Downie. 1987. HLA genetic determinants in familial MS. Tissue Antigens 29: 7-12.

Fraumeni, J. F., C. L. Vogel, and V. T. DeVita. 1969. Familial chronic lymphocytic leukemia. Annals of Internal Medicine 71: 279-284.

Froggatt, N. J., J. Koch, R. Davies, D. G. Evans, A. Clamp, W. J. Quarrell, J. Weissenbach, S. V. Hodgson, B. A. Ponder, D. E. Barton, and E. R. Maher. 1995. Genetic linkage analysis in hereditary non-polyposis colon cancer syndrome. Journal of Medical Genetics 32: 352-357.

Garay, R. P., G. Dagher, M. G. Pernollet, M. A. Devynck, and P. Meyer. 1980. Inherited defect in a Na+,K+ co-transport system in erythrocytes from essential hypertensive patients. Nature 284: 281-283.

Gardner, R. J., J. W. Hanson, V. V. Ionasescu, H. H. Ardinger, D. J. Skorton, L. T. Mahoney, and M. N. Hart. 1987. Dominantly inherited dilated cardiomyopathy. American Journal of Medical Genetics 27: 61-73.

Gill, L., M. Goldacre, H. Simmons, G. Bettley, and M. Griffith. 1993. Computerised linking of medical records: methodological guidelines. Journal of Epidemiology and Community Health 47: 316-319.

Goate, A., M. C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, L. Guiffra, A. Haynes, N. Irving, L. James, and R. Mant. 1991. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature 349: 704-706.

Goldgar, D.E., D. F. Easton, L. A. Cannon-Albright, and M. H. Skolnick. 1994. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. Journal of the National Cancer Institute 86 (November): 1600-1608.

Goldgar, D. E., P. Fields, C. M. Lewis, L. A. Cannon-Albright, G. Linker, T. Tran, and M. Skolnick. 1992. A large kindred with 17q-linked susceptibility to breast and ovarian cancer: relationship between genotype and phenotype. American Journal of Human Genetics 51: A27.

Goldgar, D.E., K. Rowe, C. M. Lewis, M. McDonald, K. Gholami, L. Cannon-Albright, and M. Skolnick. 1993. Genetic epidemiology of familial ovarian cancer in Utah. Ovarian Cancer 3: 13-21.

Greaves, S. C., A. H. Roche, J. M. Neutze, R. M. Whitlock, and A. M. Veale. 1987. Inheritance of hypertrophic cardiomyopathy: a cross sectional and M mode echocardiographic study of 50 families. British Heart Journal 58: 259-266.

Griffin, J. H., B. Evatt, T. S. Zimmerman, A. J. Kleiss, and C. Wideman. 1981. Deficiency of protein C in congential thrombotic disease. Journal of Clinical Investigation 68: 1370-1373.

Halaas, J. L., K. S. Gajiwala, M. Maffei, S. L. Cohen, B. T. Chait, D. Rabinowitz, R. L. Lallone, S. K. Burley, and J. M. Friedman. 1995. Weight-reducing effects on the plasma protein encoded by the obese gene. Science 269: 543-546.

Hasstedt, S. J., L. L. Wu, K. O. Ash, H. Kuida, and R. R. Williams. 1988. Hypertension and sodium-lithium countertransport in Utah pedigrees: evidence for major-locus inheritance. American Journal of Human Genetics 29: 14-22.

Henry, L. 1956. Anciennes familles genevoises. Paris: Presses Universitaires De France.

Hill, J. R. 1980. A survey of cancer sites by kinship in the Utah mormon population. In Banbury report 4: cancer incidence in defined populations. eds J. Cairns, J. L. Lyon, and M. Skolnick, 299-318. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Hole, B. V., and K. Wasserman. 1965. Familial emphysema. Annals of Internal Medicine 63: 1009-1017.

Houlston, R., J. Quiney, J. Mount, G. F. Watts, and B. Lewis. 1988. Lipoprotein A and coronary heart disease in familial hyperchlesterolemia. Lancet II: 405.

Howe, G. R., and J. Lindsay. 1981. A generalized iterative record linkage computer system for use in medical follw-up studies. Computers and Biomedical Research 14: 327-340.

Hsu, S. H., M. M. Chan, and W. B. Bias. 1981. Genetic control of major histocompatibility complex-linked immune responses to synthetic polypeptides in man. Proceedings of the National Academy of Science 78: 440-444.

Hunt, D., and G. Sloman. 1969. Prolapse of the posterior leaflet of the mitral valve occurring in eleven members of a family. American Heart Journal 78: 149-153.

Jaro, M.A. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of American Statistical Association 84: 414-420.

Jaro, M.A. 1993. Probabilistic linkage of large public health data files. Presented at Conference on Multi-Source Data sponsored by Center for Disease Control.

Jarvinen, H. J., J. P. Mecklin, and P. Sistonen. 1995. Screening reduces colorectal cancer rate in families with hereditary nonpolyposis colorectal cancer. Gastroenterology 108: 1405-1411.

Jeunemaitre, X., F. Soubrier, Y. V. Kotelevtsev, R. P. Lifton, C. S. Williams, A. Charru, S. C. Hunt, P. N. Hopkins, R. R. Williams, J. M. Lalouel, and Corvol, P. 1992. Molecular basis of human hypertension: role of angiotensinogen. Cell 71: 7-20.

Jones, C. T., D. J. Brock, A. M. Chancellor, C. P. Warlow, and R. J. Swingler. 1993. Cu/Zn superoxide dismutase (SOD1) mutations and sporadic amyotrophic lateral sclerosis. Lancet 342: 1050-1051.

Jorde, L. B., J. C. Carey, and R. L. White. 1995. Medical genetics. St. Louis: Mosby-Year Book, Inc.

Judson, I. R., E. Wiltshaw, and A. C. Newland. 1985. Multiple myeloma in a pair of monozygotic twins: the first reported case. British Journal of Haematology 60: 551-554.

Keen, A. J., and M. A. Knowles. 1994. Definition of two regions of deletion on chromosome 9 in carcinoma of the bladder. Oncogene 9: 2083-2088.

Knudson, R. J. 1979. Familial emphysema discovered by James Jackson, Jr. (Letter) New England Journal of Medicine 300: 374.

Kopf, A. W., L. J. Hellman, G. S. Rogers, D. F. Gross, D. S. Rigel, R. Friedman, M. Levenstein, J. Brown, F. M. Golomb, D. F. Roses, S. L. Gumport, and M. M. Mintzis. 1986. Familial malignant melanoma. Journal of the American Medical Association 256: 1915-1919.

Knuth, D. E. 1973. The art of computer programming, volume 3: sorting and searching. Reading, MA. : Addison-Wesley.

Larson, R. K., and M. L. Barman. 1965. The familial occurrence of chronic obstructive pulmonary disease. Annals of Internal Medicine 63: 1001-1008.

Lazzarini, A. M., R. H. Myers, T. R. Zimmerman, M. H. Mark, L. I. Golbe, J. I. Sage, W. G. Johnson, and R. C. Duvoisin. 1994. A clinical genetic study of Parkinson's disease: evidence for dominant transmission. Neurology 44: 499-506.

Levinson, A. K., D. E. Johnson, L. C. Strong, S. Pathak, V. Huff, and G. F. Saunders. 1990. Familial renal cell carcinoma: hereditary or coincidental?. Journal of Urology 144: 849-851.

Li, F. P., D. J. Marchetto, and R. S. Brown. 1982. Familial renal carcinoma. Cancer Genetics and Cytogenetics 7: 271-275.

Lynch, H. T., and J. F. Lynch. 1987. Breast cancer genetics: clinical nuances. In Cancer genetics in women volume I. eds H. T. Lynch and S. Kullander. Boca Raton, FL: CRC Press, Inc.

Mack, T. M., W. Cozen, D. K. Shibata, L. M. Weiss, B. N. Nathwani, and A. M. Hernandez. 1995. Concordance for Hodgkin's disease in identical twins suggesting genetic susceptibility to the young-adult form of the disease. New England Journal of Medicine 332: 413-418.

MacLennan, B. A., E. Y. Tsoi, C. Mguire, and A. A. Adgey. 1987. Familial idiopathic congestive cardiomyopathy in three generations: a family study with eight affected members. Quarterly Journal of Medicine 63: 335-347.

MacMahon, B. 1966. Epidemiology of Hodgkin's disease. Cancer Research 26: 1189-1200.

Malécot, G. 1948. Les mathematiques de l'heredite. Paris: Masson et Cie.

McWhorter, W. P., A. D. Hernandez, A. W. Meikle, D. A. Terreros, J. A. Smith, M. H. Skolnick, L. A. Cannon-Albright, and H. J. Etre. 1992. A screening study of prostate cancer in high risk families. The Journal of Urology 148 (September): 826-828.

Meikle, A. W., J. A. Smith, and D. W. West. 1985. Familial factors affecting prostastic cancer risk and plasma sex steroid levels. Prostate 6: 121-128.

Meyer, C. G., M. Gallin, K. D. Erttmann, N. Brattig, L. Schnittger, A. Gelhaus, E. Tannich, and A. B. Begovich, et al. 1994. HLA-D alleles associated with generalized disease, localized disease, and putative immunity in onchocera volvulus infection. Proceedings of the National Academy of Science 91: 7515-7519.

Miettinen, O. S. 1985. Theoretical epidemiology: principles of occurrence research in medicine. New York: John Wiley & Sons.

Miki, Y., J. Swenson, D. Shattuck-Eidens, P. A. Futreal, K. Harshman, S. Tavtigian, and Q. Liu, et al. 1994. A Strong candidate for the breast and ovarian susceptibility gene BRCA1. Science 266: 66-71.

Mineau, G. P., L. L. Bean, and M. Skolnick. 1979. Mormon demographic history. II. The family life cycle and natural fertility. Population Studies 33: 429-446.

Mineau, G. P., L. L. Bean, and D. L. Anderton. 1989. Description and evaluation of linkage of the 1880 census to familiy genealogies. Historical Methods 22: 144-157.

Merajver, S. D., T. M. Pham, R. F. Caduff, M. Chen, E. L. Poy, K. A. Cooney, B. L. Weber, F. S. Collins, C. Johnston, and T. S. Frank. 1995. Somatic mutations in the BRCA1 gene in sporadic ovarian tumours. Nature Genetics 9: 439-443.

Mueller, N., A. Evans, N. L. Harris, G. W. Comstock, E. Jellum, K. Magnus, N. Orentreich, B. F. Polk, and J. Vogelman. 1989. Hodgkin's disease and Epstein-Barr virus: altered antibody pattern before diagnosis. New England Journal of Medicine 320: 689-695.

Munoz, N., R. J. Davidson, B. Witthoff, J. E. Ericsson, and G. deThe. 1978. Infectious mononucleosis and Hodgkin's disease. International Journal of Cancer 22: 10-13.

Nakano, H., F. Yamamoto, C. Neville, D. Evans, T. Mizuno, and M. Perucho. 1984. Isolation of transforming sequences of two human lung carcinomas: structural and functional anaylsis of the activated c-kras oncogenes. Proceedings of the National Academy of Science 81: 71-75.

Narod, S., D. Ford, P. Devilee, R. B. Barkardottir, H. T. Lynch, S. A. Smith, B. A. Ponder, and B. L. Weber, et al. 1995. An evaluation of genetic heterogenity in 145 breast-ovarian families. American Journal of Human Genetics 56: 254-264.

Nesje, O. A., and K. F. Kordt. 1970. Hypoantithrombinemi som arsak til mesenterialvenethrombose. Nord. Medicine 83: 367-368.

Newcombe, H.B. 1988. Handbook of record linkage. Oxford: Oxford University Press.

Newcombe, H. B., and J. D. Abbatt. 1983. Probabilistic record linkage in epidemiology: principles employed. Ottawa,Ontario:Eldorado Resources Limited.

Newcombe, H. B., and J. M. Kennedy. 1962. Record Linkage. Communication of the association for computing machinary 5: 563-566.

Ottman, R., M. C. Pike, M. C. King, and B. E. Henderson. 1983. Practical guide for estimating risk for familial breast cancer. Lancet II: 556-558.

Paganini-Hill, M. A., M. J. Cullen, M. B. Baker, R. Hecht, D. Winters, T. Boone, and F. Collins. 1981. The S-leut anthropometric traits: genetic analysis. American Journal of Physiology and Anthropology 55: 55-67.

Pellymounter, M. A., M. J. Cullen, M. B. Baker, R. Hecht, D. Winters, T. Boone. and F. Collins. 1995. Effects of the obese gene product on body weight regulation in ob/ob mice. Science 269: 540-542.

Pramatarova, A., D. A. Figlewicz, A. Krizus, F. Y. Han, I. Ceballos-Picot, A. Nicole, M. Dib, V. Meininger, R. H. Brown, and G. A. Rouleau. 1995. Identification of new mutations in the Cu/Zn superoxide dismutase gene of patients with familial amyotrophic lateral sclerosis. American Journal of Human Genetics 56: 592-596.

Risch, A., D. M. A. Wallace, S. Bathers, and E. Sim. 1995. Slow N-acetylation genotype is a susceptibility factor in occupational and smoking related lung cancer. Human Molecular Genetics 4: 231-236.

Ronkainen, A., J. Hernesniemi, and M. Ryynanen. 1993. Familial subarachnoid hemmorrhage in East Finland, 1977-1990. Neurosurgery 33: 787-797.

Rosen, D. R., T. Siddique, D. Patterson, D. A. Figlewicz, P. Sapp, A. Hentati, D. Donaldson, J. Goto, and J. P. O'Regan, et al. 1993. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature 362: 59-62.

Ross, L. L., S. M. Sharp, and A. Wajda. 1989. Assessing data quality: a computerized approach. Social Science in Medicine 28: 175-182.

Roos, L. L., and A. Wajda. 1991. Record linking strategies. Methods of Information in Medicine 30: 117-123.

Rotter, J. I., J. Q. Sones, I. M. Samloff, C. T. Richardson, J. M. Guursky, J. H. Walsh, and D. L. Rimoin. 1979. Duodenal-ulcer disease associated with elevated serum pepsinogen I: an inherited autosomal dominant disorder. New England Journal of Medicine 300: 63-66.

Schievink, W. I., D. J. Schaid, H. M. Rogers, D. G. Piepgras, and V. V. Michels. 1994. On the inheritance of intracranial aneurysms. Stroke 25: 2028-2037.

Shell, W. E., J. A. Walton, M. E. Clifford, and P. W. Willis. 1969. The familial occurence of the syndrome of mid-late systolic click and late systolic murmur. Circulation 39: 327-337.

Skolnick, M. 1973. The resolution of ambiguities in record linkage. In Identifying people in the past, ed E. A. Wrigley, 102-127. London: Edward Arnold Press.

Skolnick, M. 1980. The Utah genealogical data base: a resource for genetic epidemiology in Banbury report no 4: cancer incidence in defined populations, 285-297. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Skolnick, M., L. L. Bean, S. M. Dintelman, and G. Mineau. 1979. A computerized family history data base system. Sociology and Social Research 63: 506-523.

Skolnick, M., D. T. Bishop, and D. Carmelli. 1981. A population-based assessment of familial cancer risk in Utah mormon genealogies. In Genes, chromosomes, and neoplasia, eds F. E. Arrighi, P. N. Rao, and E. Stubblefield, 477- 500. New York:Raven Press.

Skolnick, M. H., A. Moroni, C. Cannings, and L. L. Cavalli-Sforza. 1971. The reconstruction of genealogies from parish books. In Mathematics in the archaeological and historical Sciences, eds F. R. Hodson, D. G. Kendall, and P. Tautu, 319-334. Edinburgh: Edinburgh University Press.

Skolnick, M. 1987. Priority needs in the development of genetic epidemiology. In Environmental impacts on human health, eds S. Draggon, J. J. Cohrssen, and R. R. Morrison, 5-33. New York:Praeger Publishers.

Slattery, M. L., and R. A. Kerber. 1993. A comprehensive evaluation of family history and breast cancer risk. Journal of the American Medical Association 270 (October): 1563-1568.

Smith, M. E. 1984. Record linkage: present status and methodology. Journal of Clinical Computing 13: 52-69.

Sonninen, V., and M. L. Savontaus. 1987. Hereditary multi-infarct dementia. European Neurology 27: 209-215.

Sourander, P., and J. Walinder. 1987. Hereditary multi-infarct dementia. Acta Neuropathology 27: 247-254.

Smith, M. E. 1984. Record linkage: present status and methodology. Journal of Clinical Computing 13: 52-69.

Steinberg, G. D., B. S. Carter, T. H. Beaty, B. Childs, and P. C. Walsh. 1990. Family history and risk of prostate cancer. Prostate 17: 337-347.

Teraski, P. I., M. S. Park, G. Opelz, and A. Ting. 1976. Multiple sclerosis and high incidence of a B-lymphocyte antigen. Science 193: 1245-1247.

Third, J. L., J. Montag, M. Flynn, J. Freidel, P. Laskarzewski, and C. J. Glueck. 1984. Primary and familial hypoalphalipoproteinemia. Metabolism 33: 136-146.

Tsai, Y. C., P. W. Nichols, A. L. Hiti, Z. Williams, D. G. Skinner, and P. A. Jones. 1990. Allelic losses of chromosomes 9, 11, and 17 in human bladder cancer. Cancer Research 50: 44-47.

Ushikubi, F., M. Nakajima, M. Hirata, M. Okuma, M. Fujiwara, and S. Narumiya. 1989. Purification of the thromboxane A2 receptor from human blood platelets. Journal of Biological Chemistry 264: 16496-16501.

Weatherall, D. J 1991. The new genetics and clinical practice. Oxford: Oxford University Press.

Weaver-Feldhaus, J., N. A. Gruis, M. H. Skolnick, and A. Kamb. 1994. Localization of a putative tumor suppressor gene by using homozygous deletions in melanomas. Proceedings of the National Academy of Science 91: 7563-7567.

Weder, A. B. 1986. Red-cell lithium-sodium countertransport and renal lithium clearance in hypertension. New England Journal of Medicine 314: 198-201.

Williams, R. R., M. Skolnick, D. Carmelli, A. T. Maness, S. C. Hunt, S. J. Hasstedt, G. E. Reiber, and R. K. Jones. 1978. Utah pedigree studies: design and preliminary data for premature male CHD deaths. In Genetic analysis of common diseases: applications to predictive factors in coronary disease. New York: Alan R. Liss Inc.

Woos, L. L., and A. Wajda. 1991. Record linking strategies. Methods of Information in Medicine 30: 117-123.

Wooster, R., G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, G. Micklem, R. Barfoot, R. Hamoudi, S. Patel, C. Rice, and P. Biggs, et al. 1995. Identification of the breast cancer susceptibility gene BRCA2. Nature 378: 789-792.

Zawadzki, Z. A., Y. Aizawa, M. A. Kraj, A. R. Haradin, and B. Fisher. 1977. Familial immunopathies: report of nine families and survey of literature. Cancer 40: 2094-2101.

Zhang, Y., R. Proenca, M. Maffei, M. Barone, L. Leopold, and J. M. Friedman. 1994. Positional cloning of the mouse obese gene and its human homologue. Nature 372: 425-432.

Ziemin-Van der Poel, S., N. McCabe, H. J. Gill, R. Espinosa, Y. Patel, A. Harden, P. Rubinelli, and S. D. Smith, et al. 1991. Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias. Proceedings of the National Academy of Science 88: 10735-10739.

Zonta, L. A., S. D. Jayakar, M. Bosisio, A. Galante, and V. Pennetti. 1987. Genetic analysis of human obesity in an Italian sample. Human Hereditary 37: 129-139.