Record Linkage Methodologies

The terms 'Record Linkage' and 'Matching' are used loosely to refer to a variety of activities often undertaken within Cancer Registries. These activities can be described as belonging to one of the following three scenarios: a) Two sets of data are compared with the goal to identify all the 'best' pairs or matches between the sets - this is often called a 'one -to-one relationship' approach b) Two sets of data are compared with the goal to identify all elements of one set that match to a particular element of the second - this is often called a 'one -to-many relationship' approach c) One set of data is compared onto itself to identify elements that match with themselves - this is considered an 'unduplication' approach.

Some examples will further clarify these three scenarios. Vital status mortality linkages (often performed annually between registry master database files and mortality incidence tapes from a state or county vital statistic office) are examples of a one -to-one type approach. The intuitive premise here is that it is only possible for one person on a registry database to match with one and only one listing on the mortality tape, and conversely it is only possible for one death record to match with one and only one listing in the registry file. Geocoding linkages (routinely performed to add census tract specific data elements to a registry by matching a persons home address in the registry file to the demographic file) are examples of one-to-many linkages. While a person on the registry file will match to one and only record in the address based demographic file, it is possible for any specific address in the demographic file can match to multiple entries in the registry file - notably when multiple records within one family occur, or when multiple records within the same apartment or multi dwelling address occur. And finally the unduplication type of approach to record linkage is familiar to many, as this process is commonly required by funding or oversight agencies to insure accuracy in incidence reporting ability. Processing using this approach yields multiple (not limited to pairs) matches within the single set of data (the registry file). Although uncommon in times of large databases and fast computers, each of these scenarios are equally applicable to manual, paper based exercises as well as computerized projects. The rest of discussions in this section however concentrate on computer based record linkage systems.

Deterministic and Probabilistic Methodologies

Computerized record linkage algorithms vary widely and range from simple home grown code modules added to in-house operational data base systems, through complex custom written stand alone programs, all the way to high cost, feature laden application software suites available from software companies specializing in record linkage. Modern algorithms generally fall into two classes - Deterministic or Probabilistic depending on methodologies used in the software implementation. A bit of discussion of these two classes is presented first, then a practical comparison is offered which hopefully will be beneficial in comparing which algorithm is appropriate in a given situation. An underlying and obvious premise always drives the developments within the field of record linkage theory and technology - that is that no matter how nicely database systems are set up and maintained, there will always be missing, mis-keyed, transposed, and otherwise erroneous data elements contained in them. The methodologies have evolved to be able to handle these real world situations.

Deterministic linkage algorithms depend upon an entity relationship between the data elements (variables in our database) being compared that is static, pre-defined, and empirically based. No matter what size the data files, what values are present in the data, or now many values are missing in the data, the rules used in the match are the same.

Probabilistic linkage algorithms utilize a more fluid entity relationship between data elements that take into account various attributes of the data important for increasing likelihood (or probability) of match situations. Two key examples of attributes incorporated into probabilistic systems are error rates within any specific element, and frequency analysis of the values (states) of any specific element. The error rate associated with an element can be thought of as how often that particular element does NOT agree within a truly matched pair. A simple example might be the error rate associated with something like Social Security number. In cancer registry databases it is commonly known that the SSN is often mis-reported, a big reason for which is that the spouses' number is often used. True matched records in linkage projects, where name, address, birth date, and phone number all agree, might still have mis-matched Social Security numbers. The number of mis-matches within the files of this element (SSN) is thought of as the error rate. Probabilistic systems calculate and use this error rate to help understand which elements are more reliable than others. The second attribute always present within datasets used for comparison rests in the very intuitive concept of frequency analysis of value states. A probabilistic algorithm takes into account the 'commonness' of any given value for an element first, and uses this knowledge as a component in the matching process. The intuitive differences we all would recognize between the last names of 'Jones' and 'Kursmidlocker' illustrate this idea. A frequency analysis of both data files used in a probabilistic linkage project might find hundreds or thousands of records with last name values of 'Jones', while the number with of records with last name values of 'Kursmidlocker' would be much less. The probabilistic algorithm would 'say' that the probability of records with the 'Kursmidlocker' last name matching is much HIGHER than the probability that records with 'Jones' last name matches. Error rate assignments and frequency analysis are not typically used in deterministic linkage algorithms. These two examples are key elements of the probabilistic methodology, and have roots in the mathematical model developed originally to incorporate subtle relationships within the data files (Felligi, 1969)

A commonly held misconception about record linkage processes is that only variables that are accurate, complete, and very specific should be used in a project. What is missed with this approach is that even though an element may be incorrect or missing much of the time, the fact that TRUE AGREEMENTS sometimes are present makes a case for including the variable. One can see that within the limitations of a deterministic approach following this rule would make sense in minimizing the number and complexity of the comparisons. The flexibility afforded within the probabilistic methodology can take into account error prone, missing and more generic data values, without assigning an importance to agreements within such data elements that would incorrectly influence the actual match (truth) between records.

One key notion that the flexibility of the probabilistic algorithm adjusts to improve likelihood that true matches are identified is that the 'truth' of a reported match is derived by summing up of all the component 'weights' associated with agreements of the individual elements or variables. This summing of individual weights yields a 'Score' for the individual match. These scores can be rank listed from low to high, indicating matches of lower probability of truth to matches of higher probability of truth. This ranking is convenient and critical for those using the algorithm, and adjustment of matching parameters as well as visual review of results are based on the score. More will be discussed about visual review a bit further on in this section.

The probabilistic approach to record linkage adds mathematically defined relationships present within the specific data we use in the linkage to both improve success rates in identifying true matches, and reducing the number of incorrect matches (false positives).

Both have potentially huge implications to registry operations in areas of accuracy as well as personnel time required to manually review poorly matches sets. While it is difficult to estimate the size of these implications in a specific registry linkage project, a recent comparison between probabilistic versus deterministic approaches used in an unduplication linkage showed the probabilistic methodology found a bit over twice the number of true matches than were found with the deterministic linkage, and that the person-hours required to perform the match was considerably less with the probabilistic approach. (Chong, 1997) Cancer registries working with numerous linkage projects or with sizable databases should strongly consider the benefits of probabilistic record linkage methodologies.

Element Comparisons

A sophisticated set of comparison type operations is a key component of any type of record linkage system. These operations are applied to a given element in the matching files - and the operation determines whether or not that element is an agreement. If the element is in agreement then a positive weight is assigned, thus adding creditability that the records are true matches. To illustrate first consider a simple character comparison operator that just looks at each character of an element, and assigns agreement if all characters agree. If the element Last Name for each file being compared had "Smith" in one and "Smith" in the second, the simple character comparison operator would determine agreement was in order, and that the weight associated for agreement is assigned for this comparison. This comparison type operation is a simple one, and is often used for ad hoc, deterministic links between two files where a programmer might write a line of code that equates to "if last name on file 1 equals last name on file 2, and first name on file 1 equals first name on file 2, and street name on file 1 equals street name on file 2 - then the records match".

More complex comparisons are useful to allow for anomalies within the data. A 'smart' character comparison operator is a routine that allows for phonetic errors, letter transpositions, insertions and deletions. This is valuable to detect data entry errors. Complex numeric data comparison routines allow for individual numeral differences as well as for prorated value comparisons. For example in a numeric social security number string it may be desirable to allow for two or less numeral difference to allow for a person mistyping this error prone field. A prorated numeric comparison might be used in some types of geocoding match projects - prorated meaning that a house address number plus or minus one would still be considered a match, but anything more or less would not. Date and time comparison operators also are often used to incorporate threshold limits used to tolerate differences. There are also specialized Geocoding comparisons useful to look at address ranges and census specific block data. When these more complex comparisons operations are used in probabilistic systems their value is enhanced since weights of agreements can be adjusted to allow for partial or less successful agreements. Using the social security number example above and the prorated value comparison operator, if the two numbers were exactly the same the system might assign a weight value of 10 - if one of the characters in the social security numbers disagreed, a value of 8 might be assigned - if two disagreed, a value of 4 might be assigned, and if more than two disagreed the value would be 0. (note - these simplistic values are given as an example only).

File preparation and standardization

In a typical, two file record linkage project, the first step before any matching takes place is to prepare both data files for the process. Often the way the two files denote missing values or store date / time elements differ. It is usually desirable to prepare the files using the common conventions. Since many record linkages start with ASCII (commonly called flat or text files) the data is first extracted from the native file format the data is presented in. The extraction itself can have strange results - for example the SAS file system (commonly used in research and health related centers) exports the "." character (a period) when numeric values contained in the file are missing. The record linkage software either needs to be able to recognize this period as a missing character, or it should be converted to a 'blank' within the file, before the match is conducted.

Another valuable step often performed before record linkage projects begin is file standardization. This step involves converting values or formatting of individual variables to create new, standardized variables that are more appropriate for matching purposes. Common examples of standardization are with address information and name information. Consider these address strings: "123 Main Unit A" - "123 Main Street #A" - "123-A MAIN St". Common sense might tell us that all three are the same address - but computer systems might not. By parsing these strings, three new elements might be created - a number element containing "123", a name element containing "MAIN", and a unit element containing "A". Another use for standardization is to handle nicknames. For example often Andy is a nickname for Andrew, and Ted is a nickname for Edward. Dictionary based standardization can be useful to handle these common nuances within data files. Keep in mind that the standardization does not lose or change the values that originally were reported for a given element - rather they only add another way to consider comparisons, thus increasing chance of finding true matches.

File Blocking

File blocking is an often-misunderstood component of many record linkage methodologies (mostly Probabilistic based ones), and it warrants at least an introductory explanation here. File blocking, and it's associated requirement to perform multiple 'passes' or 'runs' of the linkage software, is used to effectively reduce the number of pairs of records that need to be compared. This is critical for files of any significant size, and when there are numerous elements being compared. A trivial project involving two files with 1000 records each would require 1 million comparisons if there were only one variable on each file. Even with high-speed computers currently available, this situation becomes infeasible. Consider routine mortality linkages in a large regional registry may be comparing 6 elements between a database file with 500,000 records against a yearly vital statistic file with 200,000 records. Without blocking these projects would be impossible.

The concept of blocking is to select some criteria to effectively split up, or 'block' the two files into smaller sub-files. A simple example might be to block on the month and day of birth on each file. For this block, ONLY records that have the exact same month and day of birth would even be eligible to be matched. One can see that this would reduce the size of the comparison task greatly. Granted - it is an obvious limitation that this block would miss many true matches - ones with either missing or incorrectly entered birth dates. For that reason, subsequent blocks need to be determined. Each iteration of the linkage software through the data is often called a 'pass'. Matches from earlier passes are removed from subsequent ones, so the process results in diminishing returns. It is not uncommon that 3 or more passes are required to insure that all true matches are identified. Blocking works best with reliable elements that have high number of possible values, and is a critical component in probabilistic linkage systems (Jaro, 1995).

Clerical Review

All computerized record linkage methodologies must allow for clerical or manual review of those matches that are determined by the software to be of questionable reliability. Both deterministic and probabilistic based systems allow for this situation. In systems using probabilistic methodologies the value assigned to the score (the sum of the matching weights of each element compared) is on a linear continuum - therefore higher scores refer to more reliable matches, and lower scores refer to less reliable ones. The matches that are candidates for clerical review can be determined by visually scanning a ranked listing of all matches. In practice, this can be a time consuming process, and is greatly facilitated by having an on-line review system. Such a system should display elements from each file grouped together for quick review. It also should allow for user input to choose between matches and non matches, an easy way to reverse decisions in the process so a wrong decision can be remedied, and a convenient way to return to the same place within the review file when a user chooses to end a session and pick up the review process at a later time.

Deciding on how detailed a clerical review process should be performed, or whether or not the labor intensive review process should be performed at all is based on the purpose and desired outcomes of the record linkage project. On one extreme, if the linkage is used as a routine production tool, a limited clerical review may be desirable. On the other extreme, a mortality record linkage on a research project database may rely on outcome so heavily that an extensive clerical review should be considered to insure completeness. When time and resources warrant, an extremely thorough review can be accomplished by retrieving paper records to augment computerized data, thus helping in the matching process.

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

Jaro, Matthew A., (1995) “Probabilistic Linkage of Large Public Health Data Files.” *Statistics in Medicine*, 14, 491-498

Chong, Nelson, (1997) “Use of Probabilistic Record Linkage Methodology for Detecting Duplicate Registrations in the Ontario Cancer Registry.” NAACCR Annual Conference, April 2, 1997, Talk Presentedp