Record Linkage Software in the Public Domain:

A Comparison of Link Plus, The Link King, and a "Basic" Deterministic Algorithm

Kevin M. Campbell, DrPH
Washington State Division of Alcohol and Substance Abuse
PO Box 45330
Olympia, Washington 98504-5330
voice: 360-725-3711
fax: 360-407-1044
email: campbkm@dshs.wa.gov

Dennis Deck, PhD
RMC Research
Portland, Oregon

Antoinette Krupski, PhD
Washington State Division of Alcohol and Substance Abuse
Olympia, Washington

Record Linkage Software in the Public Domain:

A Comparison of Link Plus, The Link King, and a "Basic" Deterministic Algorithm

Abstract

Objective: To compare the accuracy of a deterministic record linkage algorithm and two public domain software applications for record linkage (The Link King and Link Plus). *Design:* The three algorithms were used to unduplicate an administrative database containing personal identifiers for over 500,000 clients. Subsequently, a random sample of linked records was submitted to four research staff for blinded clerical review. Using reviewers' decisions as the "gold standard", sensitivity and positive predictive values (PPV) were estimated. *Results:* Optimally, sensitivity and PPV in the mid-90s could be obtained from both The Link King and Link Plus. *S*ensitivity and PPV using a basic deterministic algorithm were 79% and 98% respectively. *Conclusion:* The full feature-set of The Link King make it an attractive option for SAS users. Link Plus is a good choice for non-SAS users as long as necessary programming resources are available for processing record-pairs identified by Link Plus.

Key words: electronic patient records, probabilistic record linkage, deterministic record linkage.

Record Linkage Software in the Public Domain:

A Comparison of Link Plus, The Link King, and a basic Deterministic Algorithm

I. Introduction:

Administrative datasets containing client identifying information are often used for a variety of research and evaluation projects. The projects often require the linking of independently maintained client rosters in order to track service utilization across different systems. Unfortunately, clients may be represented with slightly different identifiers both within and across administrative datasets. The source of discrepancies include: use of nicknames and hyphenated names, misspelled names, transposed SSN digits and date fields, and changes in surname. Failure to appropriately deal with this problem may lead to incomplete linking of client records and, ultimately, introduce unnecessary error into the project. Many proprietary (often expensive) software applications have been developed to minimize errors when linking administrative datasets. This paper compares three public domain solutions.

II. Background:

Two record linkage strategies have been developed: probabilistic linkage and deterministic linkage. Detailed descriptions of probabilistic and deterministic algorithms have been previously published [1, 2, 3, 4, 5, 6, 7, 8]. Probabilistic linking is accomplished through statistical analysis of the similarity between data elements in “record pairs”. Each member of a record pair contains identifying information for a given individual. Ultimately, a formula is derived which generates a score for each record pair and cut-points to identify "definite" matches, "possible" matches, and "non matches". Some record linkage software allows the user to specify alternative cut-points. The formula incorporates weights specific to each of the data elements and scaling factors for many of the data elements. Weights reflect the relative importance of specific data elements in predicting a match. Scaling factors adjust the weights based on the frequency with which that specific data value occurs in the data being analyzed.

Deterministic linking is accomplished by establishing specific criteria about which data elements need to "match" in order to accept the link as valid. Simple deterministic protocols require each of the personal identifiers to match exactly. More complex deterministic algorithms [6, 9, 10, 11] allow some discrepancy through incorporation of "fuzzy" equivalence algorithms (e.g., phonetic equivalence). As a general rule [4, 10], the positive predictive value (PPV) of deterministic protocols are slightly higher that those of probabilistic protocols. (Positive predictive value is defined as the proportion of linked records that are valid links.) The sensitivity of deterministic protocols are usually lower than those produced by probabilistic protocols. (Sensitivity is defined as the proportion of all valid links that were captured by the linkage protocol).

Although probabilistic algorithms often use a rudimentary deterministic algorithm to bootstrap the probabilistic estimation process and the utility of integrating deterministic and probabilistic protocols has been demonstrated [12, 13], only one record linkage application (The Link King) fully integrates an intricate deterministic algorithm with a probabilistic algorithm. Records processed by The Link King are independently linked by these algorithms. A crosstabluation of the deterministic and probabilistic solutions guides the user in the selection of links.

Numerous proprietary record linkage products are readily discovered through an internet searches using such terms as “record linkage software”, “probabilistic record linkage”, and “dedupe software” but little is found to guide the selection of software. California Health Care Foundation’s (CHCF) recent report describes a number of commercial record linkage programs the low end of the cost continuum ($350 to $11,000) [14].   The CHCF, however, did not compare software performance citing a lack of a widely accepted method to evaluate record linkage accuracy

This report builds on the work of the CHCF by comparing public domain solutions to the record linkage problem. Specifically, this reports compares Link Plus (developed by the Centers for Disease Control) with The Link King (developed at Washington State’s Division of Alcohol and Substance Abuse using a probabilistic algorithm developed by MEDSTAT for the Substance Abuse and Mental Health Services Administration), and a deterministic algorithm similar to those developed by Gomatam and Carter, Weiner et al., and Grannis, Overhage and McDonald [4, 9,10]. The deterministic algorithm is referred to as the “basic” deterministic algorithm in the remainder of this paper.

In addition to The Link King and Link Plus, other public domain solutions have been identified [13, 15, 16] but are not included in this evaluation because they are either a) a series of macros requiring the skills of an advanced programmer to implement (rather than a fully developed application or easily programmed "basic" algorithm) or b) not readily available in an English version.

II. Research Questions

This inquiry compares the relative accuracy of Link Plus, The Link King, and a basic deterministic record linkage algorithm in unduplicating a large administrative dataset.

This "black box" evaluation does not include a detailed comparison of the specific algorithms used by The Link King and Link Plus. Technical details of Link Plus - required to make such a comparison – are not readily available. Technical details of The Link King's algorithms for blocking and deterministic/ probabilistic record-linkage are available to interested parties [3, 6]. A brief comparison of the software packages is provided in Appendix A.

III. Methods:

A. Unduplication of Sample Dataset

Link Plus (www.cdc.gov/cancer/registryplus/lp.htm), The Link King (www.the-link-king.com) and a basic deterministic algorithm were used to unduplicate the client database of Washington State's Division of Alcohol and Substance Abuse (DASA). DASA's client database contains over 600,000 records. Upon receipt of DASA services, clients are assigned a ClientID to facilitate linkage of services associated with that admission. Clients receiving services from multiple unrelated providers end up with multiple unrelated ClientIDs. Periodic unduplication of the client listing is necessary to identify instances where a single client is represented under multiple ClientIDs. Based on previous unduplications of DASA's administrative dataset, it is known that approximately 26% of clients in the database are represented multiple times under different ClientIDs.

Table I details criteria required for a 'basic' deterministic linkage. Each row represents criteria necessary for a deterministic match. All conditions in a given row must be met. If the conditions in any of the rows are met, the record pair is considered a deterministic match. Conditions in the first row of Table I were found to produce sensitivities of 87-88% with 100% PPV [9].

Linkages were established using subjects' first name, last name, middle name, maiden name, gender, race, birth date, and social security number. Substantial missing values were found for maiden name (93%), SSN (32%), and middle name (18%). Missing values for all other data elements were negligible.

Both The Link King and Link Plus contain controls that allow users to customize the linkage process. Based on analysis the input dataset(s) and – when necessary - user responses to yes/no questions, The Link King applies customized linkage settings. No modifications to these default settings were made in this exercise. Link Plus's on-line help system provides the user with "tips" regarding which variables to use in blocking, the most appropriate comparison protocol for various data elements, etc. These guidelines were followed to the letter. For optimal results, potential users should approach record linkage tasks with a full conceptual understanding of the process and familiarity with available settings.

Each of these three record linkage algorithms generated a listing of record pairs (i.e., pairs of client identifiers identified as potentially representing the same person). All record pairs from these programs were combined and classified to reflect the algorithm(s) generating the linkage and the relative strength of the linkage. Record pairs were classified into one of four categories based on the probabilistic score generated by Link Plus (<10, 10- 15, 16-25, 26+), into one of seven categories corresponding to The Link King's linkage "certainty" level hierarchy, and into one of two categories based on the basic deterministic algorithm (linked/not linked). In total, a cross-walk of 56 distinct levels of stratification resulted from this categorization system (4 Link Plus categories * 7 Link King Categories * 2 deterministic categories = 56 strata).

The four-level classification of linkages generated by Link Plus (i.e. <10, 10-15, 16-25, and 26+) was based on Link Plus's recommendation to set the cut-point between 10 and 15 when using the matching variables employed in this evaluation.

B. Clerical Review of Linked Records

1) Selection of Record Pairs for Clerical Review

The sampling strategy described below was developed to generate a sample large enough to provide meaningful results for this exploratory analysis but small enough to minimize time required for review. The sampling strategy over-sampled from strata where the greatest potential for uncertainty exists.

Ten record pairs were randomly sampled from strata where both The Link King and Link Plus linked the record-pair at a high certainty level (i.e. linked by The Link King at certainty levels 1 thru 3 AND by Link Plus with score of 16 or higher).

With one exception, 20 record-pairs were randomly sampled from all remaining strata containing 1,000 or more record-pairs. (The one exception was made to avoid completely excluding one of The Link King's certainty levels where all strata for that level contained less than 1,000 record-pairs). Analyses were restricted in this manner to minimize time required for manual review while maximizing generalizability to the full analytic dataset. Ultimately, 500 record-pairs were selected from 32[1] of the 56 stratum. The 32 sampled strata contained 294,214 of the 298,739 record-pairs in the full analytic dataset (98.5% of the total). Inclusion of 20 record pairs from the 24 excluded strata would more than double the number of record pairs reviewed while representing only 1.5% of the full analytic dataset.

[1] Note to reviewer, I have included 2 additional stratum in the analysis in response to a concern raised by another reviewer. Specifically, he noted that exclusion of The Link King's "Possible Twins" certainty level due to small size (741 record pairs) was not necessarily warranted.

Sampled records were weighted such that the sum of the weights equaled the “n” of the strata from which they were selected. In this manner, weighted PPV and sensitivity estimates of sampled data approximate distributions based on the complete analytic dataset. SAS’s SURVERYMEANS procedure and the associated %SMSUB macro were used to develop confidence intervals surrounding the weighted estimates.

2). Blinded Clerical Review of Sampled Record Pairs

Sampled records given to four research staff at Washington State’s Division of Alcohol and Substance Abuse for blinded review. Reviewers were asked to classify record pairs into one of five categories: The two members of the record pair are: 1) definitely not the same person, 2) probably not the same person, 3) there is not enough information to determine whether or not they are the same person, 4) probably the same person, and 5) definitely the same person.

Although the manual review process is not error free, it is the mechanism used to resolve “uncertain” linkages in other record linkage applications [2,3,16]. Additionally, manual review was referred to as the “gold standard" in California Health Care Foundation’s review of record linkage software.

A given record pair was considered a valid link if at least three of the four reviewers classified the record pair as “probably” or “definitely” the same person and none of the reviewers classified the record pair as “definitely” or “probably” NOT the same person. Remaining record pairs were considered invalid links. Alternative decision rules yielded the same general conclusions as the decision rule described above. Results presented here are on the conservative end of the spectrum. For example, one alternative decision rule considered a link valid if three of the four reviewers classified the record-pair as “probably” or “definitely” the same person regardless of the fourth reviewer’s opinion. Application of this decision rule didn’t change the overall findings. Overall PPV and sensitivity were, however, increased.

C. Determination of the Accuracy of Record Linkage Protocols

PPV and sensitivity are used as the metric to determine the relative accuracy of the three record linkage algorithms. Results of manual review serve as the “gold standard” for the calculation of PPV and sensitivity. The PPV and sensitivity of a given linkage protocol reflects the degree of correspondence between the results obtained from manual review and those obtain from the respective linkage protocol for the 500 randomly selected record-pairs.

IV. Results:

A. Accuracy of The Link King’s Linkages

Sensitivity and positive predictive values were calculated for the following aggregations of The Link King’s certainty levels: Level 1, Levels 1 and 2, Levels 1 thru 3, Levels 1 thru 4, and Levels 1 thru 6.

In Table II, the column labeled “total n” reflects the total number of record pairs linked at the associated certainty level. The column labeled “n sampled” reflects the number of manually reviewed record-pairs at the associated certainty level. The column labeled “PPV” reflects the positive predictive value for record-pairs linked at the associated certainty level. The column labeled “aggregate PPV” reflects the positive predictive value for record-pairs linked at the associated certainty level *or higher.* Similarly, the column labeled “aggregate sensitivity” reflects the sensitivity of record-pairs linked at the associated certainty level *or higher.* Numbers in parentheses represent the 95% confidence interval.

For example, consider the row “Level 4: Moderate”: Based on manual review, 81.9% of record pairs linked at The Link King’s Certainty Level 4 were validated by manual review (PPV=81.9) and 96.1% of record pairs linked at The Link King’s Certainty Level’s 1 thru 4 were validated by manual review (Aggregate PPV=96.1). Further, 96.6 of the total number manually validated links were captured among record-pairs linked at The Link King’s Certainty Level’s 1 thru 4 (Aggregate Sensitivity=96.7).

B. Accuracy of Link Plus’s Linkages

Sensitivity and predictive values positive were calculated for the following for the following aggregations of Link Plus’s probabilistic scores: 26 or higher, 16 or higher, 10 or higher.

As detailed in Table III, Link Plus's aggregate PPV declines from 94.6% using a probabilistic cut-point of 16 to 77.0% when a cut-point of 10 is used. The decline is due to the extremely low PPV of records link based on a probabilistic score of 10-15 (17.0%). This underscores the importance of reviewing a sample of links (preferably randomly generated) to determine the most appropriate cut-point. Link Plus (using a probabilistic cut-point of 16) had similar PPV to The Link King (using Certainty Level 4 as the cut-point) while Link Plus's sensitivity was slightly lower (94.1 vs. 96.6, p<.05). When The Link King's Certainty Level 3 was used as the cut-point, Link Plus had higher sensitivity (94.1 vs. 91.4, p<.05).

Post-hoc analysis of records pairs linked by Link Plus suggest that 16 was an optimal cut-point for this task: Only 7.2% of record pairs with a Link Plus score of 10-12.5 and 32.7% of record pairs with a Link Plus generated score of 12.6-15 were validated by manual review.

C. Accuracy of Basic Deterministic Algorithm

Table IV compares the accuracy of the basic deterministic algorithm to The Link King's aggregation of certainty levels 1 through 4 and Link Plus's aggregations of scores 16 or higher. The sensitivity of the basic deterministic algorithm was substantially lower than either The Link King or Link Plus (79.1% vs. 96.7% and 94.1%, p<.05). The PPV of the basic deterministic algorithm was higher than Link Plus's PPV (97.4% vs. 94.8%).

V. Discussion

A. Discussion of Results

Consistent with previous research, the basic deterministic algorithm generated the lowest sensitivity and highest PPV of the protocols compared. The sensitivity and PPV of Link Plus's and The Link King's solutions are similar to other probabilistic algorithms reported in the literature. Compared to the basic deterministic algorithm, both The Link King's integrated protocol and Link Plus's probabilistic protocol demonstrated significantly higher sensitivity with minimal sacrifice of the PPVs *when the appropriate cut-points were used*.

Although both Link Plus and The Link King are *capable* of producing extremely accurate linkage solutions, a program's potential will not be achieved if the wrong cut-point is used. For this reason it is important to provide the user with a mechanism for selecting only those links where the degree of uncertainty is acceptable to the user. The Link King classifies linked records into one of 11 categories based on a) the protocol that established the link (deterministic, probabilistic, or both) and b) the degree of uncertainty in the linked records. The user can generate random samples of linked records in each of these 11 categories and, based upon the degree of error found in the random samples, choose to include (or exclude) any of the 11 categories in the final linkage solution. Additionally, The Link King is capable of producing estimates of PPV in each of the 11 categories based on results of manual review by the user.

B. Study Limitations

This inquiry was conducted at Washington State's Division of Alcohol and Substance Abuse by the developer of The Link King. Every effort was made, however, to provide an objective inquiry. A draft of the final paper was submitted to Link Plus's developers and their comments were fully integrated into this report.

DASA data was extensively used in the development of The Link King and, therefore, it is possible that The Link King is particularly well suited to unduplicate DASA's administrative dataset. DASA data is particularly rich: missing data for ethnicity is negligible and SSNs are present for nearly 70% of clients. Other administrative data may contain large numbers of missing values or be missing some data elements in their entirety. In fact, DASA was not the sole data source used in the development of The Link King: a number of independently maintained administrative datasets from many Washington State agencies – with varying degrees of data completeness and quality - also contributed to the development of The Link King's algorithm.

Only DASA staff were used for manual review. Reviewers not associated with the agencies where the software was developed would be ideal. However, federal regulations regarding access to confidential health information prevented inclusion of non-DASA staff from reviewing personal identifiers of substance abusers. None of the DASA's reviewers were involved in the development of The Link King and, therefore, would not be able to distinguish linkages made by The Link King from those made by Link Plus or the deterministic algorithm. Further, reviewers were provided with linkages across the full spectrum of “definite” links to “definite non-links”. Reviewers were not asked to simply “confirm” linkages.

To fully overcome these potential sources of bias, replication of this analysis by other researchers in other situations would be required. Efforts to enhance generalizability are ongoing: In the late stages of The Link King’s development The State of Oregon’s Department of Human Services (DHS) was evaluating The Link King as a mechanism for consolidating client information across a broad range of administrative data systems. In the process, Oregon’s DHS staff manually reviewed thousands of record-pairs. Insights from this extensive manual review process were relayed to the developer and integrated into The Link King. Additionally, minor modifications have been made to The Link King based on feedback from users in a variety of settings.

VII. Conclusion

Ultimately, selection of record linkage software will depend on available resources. Organizations where SAS is used for data management/analysis would do well to use The Link King given its full feature set (i.e., random generation of record-pairs for validation, automatic data-driven classification of record-pairs into a linkage “certainty” hierarchy, PPV estimation). Link Plus is a viable alternative for organizations without a SAS license but with staff capable of post-processing analysis to determine an appropriate probabilistic score cut-point to isolate valid links.

"Ease of use" is also an important consideration in the selection of record linkage software. Fortunately, The Link King and Link plus are freely available for potential users to evaluate. Additionally, The Link King's website contains an 8-minute demonstration video (http://www.the-link-king.com/flash2.html). The video walks the user through the process of unduplicating a dataset, familiarizing the viewer with The Link King's interface. Although both The Link King and Link Plus greatly simplify the record linkage process, users should develop a general understanding of the steps involved in record linkage to insure appropriate decisions are made when setting up the linkage job. Both Link Plus and The Link King have extensive on-line help. A detailed user manual is also available for The Link King. A particularly well written description of the technical aspects of probablisitic record linkage can be found Whalen et al.'s technical monograph [3] describing the use of probabilistic matching protocols in the Substance Abuse and Mental Health Services Administration's Integrated Database Project.

Proprietary record-linkage programs are also an option although little empirical evidence establishing the accuracy of their linkage solutions is currently available.

VII. Bibliography

1. Gill L, Goldacre M, Simmons H, Bettley G, Griffith M .Computerized linking of medical records: methodological guidelines. J Epidemiol Community Health 1993; 47: 316-19.

2. Jaro M. Probabilistic linkage of large public health data files. Stat Med 1995; 14: 491-98.

3. Whalen D, Pepitone A, Graver L, Busch J. Linking Client Records from Substance Abuse, Mental Health, and Medicaid State Agencies. Rockville (MD): Substance Abuse and Mental Health Services Administration; 2001.

4. Gomatam S,Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Stat Med 2002; 21: 1485-96.

5. Clark D. Practical introduction to record linkage for injury research Inj Prev 2004; 10(3): 186-91.

6. Campbell K, Deck D, Cox C, Broderick C. The Link King User Manual [Online]. 2005 [Cited Nov 3, 2006]; Available from: URL: www.the-link-king.com\user_manual.zip.

7. Christen P, Goiser K. Quality and Complexity Measures for Data Linkage and Deduplication [Online]. 2006 [Cited Nov 3, 2006]; Available from: URL: http://cs.anu.edu.au/people/Peter.Christen/publications/qmdm-linkage.pdf

8. Newcombe H, Kennedy J, Axford S, James A. Automatic Linkage of Vital Records. Science 1959; 130: 954-959.

9. Weiner M, Stump T, Callahan C, Lewis J, McDonald C. A practical method of linking data from Medicare claims and a comprehensive electronic medical records system Int J Med Inform 2003; 71(1): 57-69.

10. Grannis S, Overhage J, McDonald C. Analysis of identifier performance using a deterministic linkage algorithm. Proceedings of American Medical Informatics Association Symposium; 2002; Philadelphia, PA. Hanley and Belfus.

11. Gomatam S, Carter R., A Computerized Stepwise Deterministic Strategy for Record Linkage. University of Florida Technical Report 615; 1999.

12. Kendrick S, Douglas M, Gardner D, Hucker D. “Best-Link Matching of Scottish Health Data Sets. Methods Inf Med 1998; 37(1): 64-68.

13. Wajda , Roos L, Layefsky M, Singleton J. Record linkage strategies: Part II. Portable software and deterministic matching. Methods Inf Med 1991; 30: 210-14.

14. Jones L, Sujansky W Patient Data Matching Software: A Buyers Guide for the Budget Conscious. California Health Care Foundation; 2004.

15. Contiero P, Tittarelli A, Tagliabue G, Maghini A, Fabiano S, Crosignani P, Tessandori R. The EpiLink record linkage software. Methods Inf Med 2005; 44(1): 66-71.

16. Dal Maso L, Braga C, Franceschi S. Methodology used for Software for Automated Linkage in Italy (SALI). J Biomed Inform 2001; 34: 387-395.

Table I: Basic Deterministic Match Criteria

Name						
Last	**First**	**Middle**	**Birth date**	**SSN**	**Sex**	**Race**
	NYIIS[*]		Partial: Month only	Exact	Exact	
NYIIS	NYIIS		Exact		Exact	Exact
NYIIS	NYIIS		Partial: Month & day	Exact		
NYIIS	NYIIS		Partial: Month & year	Exact		
NYIIS	NYIIS		Partial: Day & year	Exact		
NYIIS	NYIIS	Partial: Initial Only		Exact		
NYIIS	NYIIS		Exact	Partial: 7 digits[+]		
NYIIS		NYIIS	Exact		Exact	Exact
	NYIIS	NYIIS	Exact		Exact	Exact
	NYIIS		Exact	Exact		

[*] Name coding according to New York State Identification Information System phonetic equivalence algorithm must match exactly.

[+] Requires 7 of the 9 SSN digits to be positionally correct.

Table II

Sensitivity and Positive Predictive Value

for The Link King

		n		Aggregate	Aggregate
Certainty Level	**total n**	**sampled**	**PPV**	**PPV**[**]	**Sensitivity**[*]
Level 1:	190,476	120	97.6%	97.6%	82.8%
Highest			(96.5, 98.8)	(96.5, 98.8)	(81.7, 84.0)
Level 2:	17,063	80	95.3%	97.4%	89.8%
Very High			(88.2, 100)	(96.2, 98.6)	(88.7 90.9)
Level 3:	4,704	60	82.2%	97.1%	91.4%
High			(72.0, 92.4)	(95.9, 98.3)	(90.3 92.5)
Level 4:	16,946	80	81.9%	96.1%	96.6%
Moderate			(72.6, 91.3)	(94.9, 97.4)	(95.6 97.6)
Level 5:	741	40	25.5%	96.0%	96.7
Possible Twins			(12.1, 39.0)	(94.7, 97.3)	(95.7, 97.7)
Level 6:	3,711	40	66.5%	95.5%	97.8%
Low			(50.4, 82.6)	(94.3, 96.8)	(96.8, 98.8)

[*] Summarizes sensitivity for record pairs linked at the specified level or higher.

[**] Summarizes PPV for record pairs linked at the specified level or higher.

Table III

Sensitivity and Positive Predictive Value

for Link Plus

Probabilistic		**n**		**Aggregate**	**Aggregate**
Score	**total n**	**sampled**	**PPV**	**PPV****	**Sensitivity***
26+	132,880	100	98.6%	98.7%	58.4%
			(97.3, 99.9)	(97.4, 100.0)	(57.5, 59.3)
16-25	92,290	180	88.7%	94.6%	94.1%
			(85.5, 91.9)	(93.1, 96.1)	(93.4, 94.7)
10-15	67,636	140	17.0%	77.0%	99.0%
			(14.7, 19.3)	(75.7, 78.3)	(98.8, 99.2)

*Summarizes sensitivity for record pairs linked at the specified level or higher.

**Summarizes PPV for record pairs linked at the specified level or higher.

Table IV

Sensitivity and Positive Predictive Value

A Comparison of Basic Deterministic

to The Link King and Link Plus

		n		
method	**total n**	**sampled**	**PPV**	**Sensitivity**
Basic Deterministic	183,219	220	97.4%	79.1%
			(96.5, 98.4)	(77.9, 80.2)
The Link King	229,189	340	96.1%	96.7%
Levels 1 thru 4			(94.9, 97.4)	(95.7, 97.7)
Link Plus	225,170	280	94.8%	94.1%
Score of 16+			(93.3, 96.3)	(93.4, 94.7)

Appendix A: Comparison of Link Plus and The Link King Software

While both the Link Plus (www.cdc.gov/cancer/registryplus/lp.htm) and The Link King (www.the-link-king.com) run on a Microsoft Windows based PC, the Link King requires a base SAS license. Link Plus is a stand-alone application. This is an attractive feature given SAS's annual individual license fee of approximately $2,000. Link Plus also has greater flexibility in the **variables used for linking**, allowing up to 15 variables to be specified including user-defined variables. The Link King allows 7 pre-defined variables (first, middle, last, and maiden names as well as SSN, race, and birth date) and one user-defined variable. Both Link Plus and The Link King support a variety of **formats for the input dataset** including delimited files, MS Access data tables, and Excel spreadsheets.

The Link King is more of a fully self-contained application than Link Plus. An interface for **manual review of uncertain links** and a tool for generating **random samples of linked records for validation** are fully integrated into The Link King. An upcoming release of Link Plus will contain an interface for manual review of uncertain links.

Additionally, The Link King consolidates all records believed to represent the same person under a common "unique id" while the current version of Link Plus only provides the user with a listing of record-pairs. This feature is particularly useful when one-to-many or many-to-many linkages are expected. The Link King's consolidation includes records that were directly linked to each other and, in some cases, records that were indirectly linked together. The process of gathering indirect links into a consolidation is called “chaining”. The Link King selectively chains records in the construction of the final linkage map. The Link King allows for the possibility of disagreement with the user and empowers the user to easily modify any group of consolidated records with “point and click” functionality.

The Link King and Link Plus use similar **comparison protocols** for determining the degree of similarity between data elements. Scaling factors adjust weights for matching values, based on the relative frequencies of values being compared.   Name comparisons consider partial matches and typographical errors, misspellings, and hyphenated names and the occurrence of the middle initial only versus the full middle name. The Link King utilizes three phonetic equivalence algorithms (Double Metaphone, Soundex, and the New York State Identification Information System) while Link Plus only utilizes the Soundex algorithm).   The Link King’s user manual details how The Link King incorporates phonetic comparison protocols (page 18. and Appendix D) into the record-linkage algorithms (6). Date and SSN comparisons in both applications also consider partial matching, accounting for typographical errors and transposition of digits. Both applications allow user-defined values to be treated as missing data. In addition to these common protocols, The Link King features a **nickname look-up-table** and **gender imputation**.

Both programs allow the user to adjust **blocking criteria** although in different manners. Link Plus allows the user to specify up to five blocking variables.   The blocked dataset will consist of all records-pairs where any one of the specified blocking variables to match. Link Plus recommends, at a minimum, the following blocking variables: SOUNDEX version of Last Name, Social Security Number, and birth date. In contrast, The Link King allows the user to select from one of three "Blocking Levels" (low, medium, and high). The Link King's "low" setting is a modified version of criteria developed by MEDSTAT for the Substance Abuse and Mental Health Administration's Integrated Database Project. According to MEDSTAT criteria, the blocked dataset would include record pairs meeting any one of the following criteria: a) SSN matched, b) date of birth and phonetic (NYSIIS) last name matched, c) date of birth, gender, and phonetic first name matched, or d) gender, phonetic first name, and phonetic last name matched. The Link King's "medium" and "high" blocking levels incrementally expand the blocking criteria to include more records in the decision space. Appendix C in The Link King's user manual details blocking criteria for each of the three blocking levels available (6).