Using Test Databases to Evaluate Record Linkage Models and Train Linkage Practitioners

Michael H. McGlincy
Strategic Matching, Inc.
PO Box 334, Morrisonville, NY 12962
Phone 518 643 8485, mcglincym@strategicmatching.com

Abstract

Traffic safety research grantees develop Crash Outcome Data Evaluation Systems (CODES) by linking police records to medical records. The process is complex because most records lack unique identifiers and exhibit high levels of misclassification and nonresponse. Grantees learn to develop Bayesian models which compare quasi-identifiers in order to estimate the probability that a record pair is a true link. To help teach effective modeling techniques, the CODES program uses test databases for which true link status is known. A data generator creates test databases for each grantee by simulating crash and medical events. Quasi-identifiers specific to each grantee are drawn from multinomial distributions. Case duplication, item misclassification, and item nonresponse are simulated as Bernoulli events. The generator captures clustering of data for vehicle occupants that occurs in real life.

Keywords: Bayesian Record Linkage Simulation

1. Crash Outcome Data Evaluation Systems

The National Highway Traffic Safety Administration (NHTSA) provides grants to traffic safety researchers to develop and analyze Crash Outcome Data Evaluation Systems (CODES). These state-specific systems link police crash reports to medical treatment records, typically ambulance run reports from Emergency Medical Services, emergency department treatment records from hospitals, or inpatient discharge records from hospitals (Runge, 2000). Files to be linked typically contain between several thousand and a few million records. CODES grantees receive commercial record linkage software (CODES2000 or LINKSOLV) and training for its use from the author (McGlincy, 2004).

CODES researchers develop record linkage models, create linked datasets, and conduct outcome studies. In CODES linkage models, true link status is treated as a latent variable, missing on all candidate record pairs. Bayesian analysis is used to estimate the posterior probability that each candidate record pair is a true link given comparison outcomes for all specified quasi-identifiers. Quasi-identifiers are included for both events and persons because one event can involve many persons and one person can be involved in many events. Multiple imputation techniques (Schafer, 1997) are used to draw and analyze multiple complete sets of linked record pairs. Disposition of candidate pairs is determined completely automatically through imputation, not through clerical review of uncertain pairs. For example, each candidate pair with posterior probability of 0.5 is randomly selected as a true link in about half of all imputations. Elements of Bayesian record linkage for CODES are summarized in Section 2.

The process is complex but necessary in order to obtain an unbiased representation of the unknown set of true linked pairs from all CODES states. Most of the administrative datasets available to CODES researchers lack unique identifiers. Many quasi-identifiers available for the linkage process exhibit high levels of nonresponse and misclassification (Greenberg, 1996). Different quasi-identifiers are available in each state, with different data characteristics. In some states, some quasi-identifiers are missing by design on crash reports for certain sub-populations. More information is collected about drivers than about passengers or more information about injured occupants than about uninjured (as determined on scene by police). In all states, records about rare situations such as 80-year-old drivers link with higher posterior probabilities than records about more common situations such as 18-year-old drivers. As consequences of these data characteristics, many true matched pairs have low posterior probabilities, many true unmatched pairs have high probabilities, and sets consisting of only high-probability pairs are likely to be biased as well as incomplete.

Collectively, multiple linkage imputations can provide unbiased estimates about the unknown set of true linked pairs. They also ensure that uncertainty caused by low-probability true links and high-probability false links is reflected in outcome studies. All CODES states can produce unbiased study results using multiple imputation techniques. States with strong quasi-identifiers will have low between-imputation variances while states with weak identifiers will have high variances. No single imputation of true link status has these properties, such as might be obtained by using only maximum likelihood estimates.

Understanding and correctly applying all of the concepts of Bayesian record linkage are challenging tasks. New linkages for CODES are usually done once each year by each grantee after new annual datasets become available. For many grantees, this annual linkage project is their only opportunity to develop and apply linkage expertise with real data. Furthermore, many CODES researchers are primarily traffic safety experts rather than record linkage experts. Consequently, NHTSA provides training sessions so that grantees can improve their skills for developing and validating record linkage models.

Choosing data for effective hands-on training has been problematic. When each state team trains with its own data then important general principles can be obscured by state-specific questions and difficulties. Most importantly, true link status is unknown for real datasets so that trainees cannot easily distinguish correct results from mistakes. To help improve training effectiveness, the capability to create artificial test datasets for which true link status is known for every record pair was added to the CODES2000 and LINKSOLV software. Even so, if each state team were to train with the same artificial test data then many state-specific questions could remain unanswered. Consequently, data simulation algorithms were designed so that artificial test datasets could be tailored to match real datasets for each state in many respects. The process for creating artificial data for CODES training is described in Section 3.

Three training exercises become straightforward when using artificial test datasets. First, trainees can examine posterior probabilities for all true links to improve their understanding of Bayesian record linkage results. Second, trainees can measure the goodness of fit of a linkage model to test whether calculated posterior probabilities are accurate. Third, trainees can compare goodness of fit measures for alternative models to help contrast them. Results from hypothetical training exercises are described in Section 4.

Section 5 describes open issues related to using artificial data for record linkage training.

2. Bayesian Record Linkage

This section summarizes key elements of Bayesian record linkage for CODES (McGlincy, 2004). CODES record linkage is similar to analyses using mixture models described by Larsen (2004), Winkler (1988, 1989, 1993, 1994), and others but implementation differs in several important details. Posterior odds for a true link are obtained by applying Bayes' rule for odds (Gelman *et al*., 2004, pg. 9)

$$\text{Posterior Odds} = \text{Prior Odds} \times \text{Likelihood Ratio}.$$

Prior odds are strongly informative, set equal to the number of true links (or matched pairs) divided by the number of true non-links (or unmatched pairs), based on reported information. For example, police might note on their crash reports whenever an injured crash victim is transported by ambulance. Or, EMS teams might note on their ambulance run reports whenever they respond to a motor vehicle crash.

The likelihood ratio is set equal to the test statistic m/u defined by Fellegi and Sunter (1969), where m is the conditional probability of observing any comparison outcome on a true matched pair and u is the conditional probability of observing the same comparison outcome on a true unmatched pair. The ratio is calculated as the product of m_k / u_k for each comparison field k,

$$m / u = \Pi_k \, m_k / u_k.$$

Each comparison field (quasi-identifier) can agree on a specific value, disagree, or be missing on either file. Each m_k and u_k, is calculated using formulas given by Fellegi and Sunter as their Method I. For computational convenience, some calculations are done using match weights, defined as

$$\text{Match Weight} = \log_2 (m / u) = \Sigma_k \log_2 (m_k / u_k).$$

Method I formulas apply when observations are independently drawn from identical distributions, comparison outcomes are independent, data values have known multinomial distributions, probabilities of item nonresponse are known for all quasi-identifiers and are independent of data values, and probabilities of misclassification are known for all quasi-identifiers and are independent of data values. Real datasets seldom satisfy all of the conditions for applying their formulas exactly as presented by Fellegi and Sunter. However, their Method I is still an appropriate starting model for CODES linkages because it incorporates nonresponse and misclassification explicitly, common occurrences in most CODES datasets.

The main difficulty with using Method I for m/u is that parameters needed to calculate each m_k and u_k, are not known. Suppose records to be linked are from populations A and B. Multinomial distributions for quasi-identifiers in A or B can be estimated from reported values, as can probabilities of nonresponse. Multinomial distributions for quasi-identifiers in true matched pairs ($A \cap B$) and probabilities of misclassification cannot. However, the latter parameters can be estimated from the set of true matched pairs. For CODES, model parameters and true link status are drawn simultaneously from their joint probability distribution through Markov Chain Monte

Carlo data augmentation (McGlincy, 2004; Schafer, 1997). Linkage practitioners supply starting values for all model parameters. Given the parameters, the missing true link status is imputed for all candidate pairs using Bayes' rule for odds and Method I for *m/u*. Given the true link status, new values for model parameters are drawn from their posterior distributions. These are the Imputation or I-step and Posterior or P-step, respectively (Schafer, 1997, pg. 72). The steps are iterated until stationarity (convergence in distribution) is achieved. Each linkage imputation is drawn from an independent chain.

Another difficulty with using Method I for *m/u* is that data for vehicle occupants are not independent, identically distributed observations. Many crashes involve more than one vehicle and many vehicles have more than one occupant. Everyone in a crash is injured at the same time and place. All crash victims are transported to the same hospitals near crash locations. In practice, CODES researchers apply this heuristic: Method I can be used but the amount of information contributed by event quasi-identifiers must be limited. For example, if there were two occupants in each crash then event quasi-identifiers alone should give posterior probabilities no greater than 0.5.

3. Creating Artificial Data

3.1 Preparation for Simulation

The data generator program creates artificial test databases suitable for each grantee by simulating the occurrence and documentation of motor vehicle crash events, other types of injury events, and related medical treatment events. The data generation process captures clustering of data for vehicle occupants that occurs in real life.

Monte Carlo simulation algorithms are used to draw a full set of typical event-related and person-related quasi-identifiers from simplified multinomial distributions loosely based on real data. To create state-specific artificial data, a linkage practitioner specifies the simulated duration in days and the average number of various classes of injury events per day based on state experience. The practitioner also specifies probabilities for item misclassification, item nonresponse, and case duplication based on state experience.

3.2 Simulate State-Specific Locations

State-specific locations (zip codes, cities, towns, and counties) are selected from a master list. The master list is created from information about real zip codes from the United States Postal Service (USPS). State-specific acute care hospitals are selected from a master list. The master list is created from information about real hospital facilities from the Centers for Medicare and Medicaid Services (CMS). The hospital closest to each zip code is determined by straight-line distance.

3.3 Simulate Date, Time, and Location of Events

The total number of injury events for each class is set to state average experience times the number of days simulated. Date and time of each event are randomly assigned within the specified horizon. Location of each event is randomly drawn from the state-specific list of locations. This places more events in urban areas with many zip codes than in rural areas with one zip code.

3.4 Simulate Crash Events

For each crash event, the type of collision and the number and types of vehicles involved are drawn from multinomial distributions loosely based on real data. For example, the artificial crash data includes only 4 common vehicle body types compared to 20 or more types in a typical real crash dataset. For each vehicle, the number of occupants is drawn from a multinomial distribution and a separate record is created for each person. Seating positions are fixed: the first occupant is always a driver, the second occupant is always a passenger in the right-front seat, etc. For each occupant, the use of safety equipment and injury severity are drawn from dependent multinomial distributions.

3.5 Simulate EMS Responses

Response by an EMS agency is simulated for each person injured in a crash. EMS responses are also simulated for other injured persons based on state experience. The delay from injury occurrence to EMS call is drawn from a normal distribution loosely based on real data. The delay from EMS call to arrival on scene is drawn from a normal distribution. EMS action (transport or treat only) is randomly drawn from a binomial distribution. Transports are simulated to the nearest hospital and transport time is drawn from a normal distribution.

3.6 Simulate Medical Treatment

Treatment at an acute care facility is simulated for each injured person transported by EMS. Treatment is also simulated for injured persons arriving at an emergency department by other means, based on state experience. The delay from arrival to start of treatment is drawn from a normal distribution. Treatment duration is drawn from a log normal distribution. The delay from start of treatment to admission as an inpatient is drawn from a normal distribution.

3.7 Simulate Personal Quasi-Identifiers

Personal quasi-identifiers are randomly drawn from multinomial distributions for each vehicle occupant, injured or uninjured, and each person injured in other events. True unique identifiers are added to all records so that true link status of any record pair can be determined by inspection.

3.8 Simulate Common Data Errors

Three common data collection problems are simulated. Case duplication, item misclassification, and item nonresponse are simulated as Bernoulli events with probabilities of occurrence set to match state experience. Cases are duplicated before misclassification and nonresponse are introduced so that duplicate records receive independent errors. Thus, true duplicates may not be exact duplicates.

4. Training with Artificial Data

4.1 Hypothetical Training Exercise

Consider a hypothetical training exercise for Nebraska researchers. First, a trainee specifies that artificial police crash reports and EMS ambulance run reports should be created for Nebraska for a 31-day period. Crashes are specified to occur at an average rate of 107 per day. EMS ambulance runs for other injury events are specified to occur at an average rate of 44 per day. Probabilities of misreporting and nonresponse are set to 0.02 and 0.05, respectively, for each person-related and event-related quasi-identifier. Given these specifications, the simulation program creates test datasets containing artificial police crash report records for 7,701 people in 3,317 crashes and EMS run report records for 3,494 injured persons. There are 2,130 true links (same person and event) between the crash and EMS datasets.

4.2 Starting Linkage Model

Trainees specify a starting linkage model. The model tested here compares the seven quasi-identifiers shown in Table 1.

Table 1. Quasi-Identifiers in Crash and EMS Datasets for Starting Linkage Model

ID Type	Crash Identifier	EMS Identifier	Information (Bits)
Event	Crash Date	Call Date	4.9
Event	Crash Zip	Incident Zip	9.1
Event	Collision Type	Collision Type	1.7
Person	Age	Age	6.2
Person	Birthday	Birthday	8.5
Person	Seat Position	Seat Position	0.9
Person	EMS Action	Disposition	1.0

Three quasi-identifiers are event-specific identifiers and four are person-specific identifiers. Information entropy is shown as a measure of the average contribution from each quasi-identifier (Goldman, 1953).

The model includes two match passes: pass 1 considers candidate record pairs which agree on person age and pass 2 considers pairs which agree on county of event. Passes are run independently with identical parameters and results are merged. Only pairs with posterior probabilities greater than 0.01 (a low but arbitrary *cutoff probability*) are tabulated. Five parallel Markov chains are used to determine posterior probabilities and five independent linkage imputations are drawn.

Trainees measure goodness of fit of linkage models as for logistic regression models (Hosmer and Lemeshow, 2000):

1. Sort all tabulated candidate pairs by posterior probability and divide the table into deciles.
2. Determine the actual number of true links in each decile by inspection.
3. Determine the expected number of true links in each decile by summing posterior probabilities.
4. Determine a *chi square* test statistic by comparing actual counts to expected counts.

The analysis is repeated for each imputation and results are combined using standard methods (Little and Rubin, 2002). Cell counts are averaged and the repeated-imputation p value is $\Pr(F_{k,b} > F_c)$, where k=10 degrees of freedom, and b and F_c are derived from the five test statistics. Total expected true links are not constrained to equal total actual true links.

Table 2. Goodness of Fit of Starting Linkage Model P Value < 0.01

Decile	Pairs In Decile	Actual True Links	Expected True Links
1	312	2.0	5.2
2	312	5.0	14.1
3	312	25.4	53.5
4	312	185.0	252.4
5	312	307.0	311.4
6	312	311.6	312.0
7	312	312.0	312.0
8	312	312.0	312.0
9	312	312.0	312.0
10	312	312.0	312.0
Totals	3,120	2,084.0	2,196.6

The fit is poor (p value < 0.01) for the starting model. By inspecting Table 2, trainees learn that the model produces many false positives in the lower deciles, with about 5% excess links overall. The model does not find all true links. There are 2,130 – 2,084 = 44 missing links (false negatives).

4.3 Improved Linkage Model

The fit is poor for the starting model because the model includes too much event information. The test crash dataset includes 7,701 people in 3,317 crashes. This means that event quasi-identifiers alone can only determine the correct link, on average, with probability 3,317 / 7,701 = 0.43. For the test datasets used here, probability = 0.43 is obtained with match weight = 13.2. On average, the three event quasi-identifiers contribute match weight = 4.9 + 9.1 + 1.7 = 15.7. Trainees improve the model by reducing the contribution from event quasi-identifiers. For example, one could drop event date, replace event zip code with county, or reduce calculated weights by an appropriate factor. For the results in Table 3, event weights were multiplied by 13.2 / 15.7 = 0.84.

Table 3. Goodness of Fit of Improved Linkage Model P Value > 0.99

Decile	Pairs In Decile	Actual True Links	Expected True Links
1	235	8.8	6.6
2	236	132.6	136.0
3	236	231.4	233.2
4	236	235.2	235.7
5	236	236.0	236.0
6	235	235.0	235.0
7	236	236.0	236.0
8	236	236.0	236.0
9	236	236.0	236.0
10	236	236.0	236.0
Totals	2,358	2,023.0	2,026.5

By inspecting Table 3, trainees learn that model fit is much improved (*p* value > 0.99) after correcting for excess event information. The average posterior probability for actual true links is 0.96. The model does not find all true links. There are 2,130 – 2,023 = 107 missing links (false negatives). Trainees investigate possible ways to correct the false negatives. Some false negatives could be corrected by adding match passes. Others could be corrected by adding or changing quasi-identifiers.

4.4 Reduced Linkage Model

Trainees investigate the effects of using different sets of quasi-identifiers. The reduced linkage model tested here includes five quasi-identifiers as shown in Table 4. Many CODES datasets do not include dates of birth. Lack of this important quasi-identifier makes linkage results more uncertain.

Table 4. Quasi-Identifiers in Crash and EMS Datasets for Reduced Linkage Model

ID Type	Crash Identifier	EMS Identifier	Information (Bits)
Event	Crash Date	Call Date	4.9
Event	Crash Zip	Incident Zip	9.1
Person	Age	Age	6.2
Person	Seat Position	Seat Position	0.9
Person	EMS Action	Disposition	1.0

Without correction, the reduced model still includes too much event information (4.9 + 9.1 = 14.0). For the results in Table 5, match weights for both event quasi-identifiers were multiplied by 13.2 / 14.0 = 0.94.

Table 5. Goodness of Fit of Reduced Linkage Model P Value = 0.80

Decile	Pairs In Decile	Actual True Links	Expected True Links
1	514	9.2	4.7
2	514	11.8	9.0
3	515	17.6	13.7
4	514	25.6	20.3
5	515	38.0	39.6
6	514	99.0	110.7
7	514	321.0	339.8
8	515	492.2	503.7
9	514	510.0	512.6
10	515	511.6	514.6
Totals	5,144	2,036.0	2,068.7

By inspecting Table 5, trainees learn that the reduced model also fits the data well (*p* value = 0.80) but that it places over twice as many candidate pairs above the cutoff probability. One cannot be as certain which links are probably true and which links are probably false with less evidence on which to base posterior probabilities. The average posterior probability for actual true links is now 0.84. The model does not find all true links. There are 2,130 – 2,036 = 74 missing links (false negatives).

The reduced model introduces more between-imputation variance into typical study results than the improved model. As one example, the sample variance of the square roots of five by-imputation *chi square* statistics is used in estimating combined *p* values shown in Tables 3 and 5. The variance is 0.3 for the improved model (Table 3) compared to 0.8 for the reduced model (Table 5).

5. Issues

5.1 Do Lessons Learned Apply to Real Data?

Training with artificial test datasets can be counter-productive unless most lessons learned apply to real datasets (Winkler, 2005). Clearly, the lessons described in Section 4 apply to many real CODES record linkage situations. Trainees learn that Bayesian posterior probabilities estimated using the CODES record linkage methodology can accurately reflect true link status. A good linkage model developed for realistic artificial test datasets is likely to be a good starting model for linking real datasets. Lessons learned linking artificial datasets are likely to reduce trial and error methods for arriving at effective linkage models.

5.2 Are Quasi-Identifiers Realistic Enough?

The collection of artificial quasi-identifiers produced by the simulation program has been expanded and modified based on informal feedback after training sessions. Most CODES states can now mimic most of their important match variables with artificial quasi-identifiers. A more systematic effort is needed to identify shortcomings, test their effects, and develop solutions. For example, simulated locations are drawn from USPS zip codes. Omaha contains about 8% of the zip codes in Nebraska but about 23% of the people. Real people (and real crashes) are more concentrated in urban areas than reflected in the artificial quasi-identifiers, crash zip and home zip.

Values of some quasi-identifiers which differ by less than a specified tolerance (using some appropriate metric) are considered agreements. For example, two event times which differ by 15 minutes or locations which differ by 5 miles might be considered agreements. Simulated events have realistic proximities in time and place but do not include relatively rare events such as helicopter transports to distant hospitals. Trainees can use simulated data to evaluate different models with different comparison tolerances for event proximity.

Quasi-identifiers created by the simulation program have fewer dependencies than real quasi-identifiers. For example, simulated crash times and locations are independent but real crash times and locations might be dependent because of rush-hour traffic patterns. Dependent quasi-identifiers cause dependent comparison outcomes (agreements on unmatched pairs), a violation of model assumptions. Trainees cannot yet use simulated data to test all realistically dependent agreements for statistical significance (*chi square p-values*), measure their strengths (*symmetric uncertainty coefficients*), and correct the linkage models.

5.3 Are Types of Errors Realistic Enough?

All misreporting is simulated as misclassification. True values are replaced with other values drawn at random from the same multinomial distribution. Some real errors are not misclassifications. For example, a person’s name might be reported with two letters transposed. As with event proximity, real quasi-identifiers which differ by less than a specified tolerance (using some appropriate metric) might be considered agreements. Trainees cannot yet use simulated data to evaluate different models with different comparison tolerances for such misreporting.

CODES linkage practitioners investigate and document types of errors and their frequencies for each quasi-identifier when preparing real data for linkage. Complicated types of errors have been found to occur for person names and street addresses which violate model assumptions. For example, common or short names might have fewer errors than uncommon or long names. Both linkage models and simulation models assume errors are equally likely for all data values. This assumption seems more plausible for quasi-identifiers in the models described here, and many CODES datasets do not include names or addresses.

Errors created by the simulation program are independent but errors in real data might be dependent. For example, age and birthday might have dependent errors if both are derived from date of birth. Dependent errors cause dependent comparison outcomes (disagreements on matched pairs), a violation of model assumptions. Trainees cannot yet use simulated data to test all realistically dependent disagreements for statistical significance (*chi square p-values*), measure their strengths (*symmetric uncertainty coefficients*), and correct the linkage models.

5.4 Are Training Times Practical?

It can take several days or more to prepare realistic artificial test datasets and run a linkage project. For example, the linkage software evaluates about 10,000,000 candidate pairs per hour. Additional time is needed to test model fit and find improvements. Such time commitments compete with preparing real data for linkage or validating real linkage results. Training with one year’s artificial data can be done but run times will be as long as for the real linkage. Training with one month’s artificial data, as described here, is much quicker but linkage results will not be as realistic.

Acknowledgements

This research was funded in part by the Governors Highway Safety Association and the National Highway Traffic Safety Administration.

References

Fellegi, I.P. and Sunter, A.B. (1969), "A theory for record linkage," *Journal of the American Statistical Association*, **64**, 1183–1210.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004), *Bayesian Data Analysis*, Chapman & Hall.

Goldman, S. (1953), *Information Theory*, Dover.

Greenberg, L. (1996). *Police Accident Report (PAR) Quality Assessment Project*. Technical Report DOT HS 808 487, National Highway Traffic Safety Administration.

Hosmer, D.W. and Lemeshow, S. (2000), *Applied Logistic Regression* (2nd Edition), Wiley.

Larsen, M.D. (2004), "Record linkage using finite mixture models," *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* Edited by A. Gelman and X.L. Meng, Wiley: 309–318.

Little, R.J.A. and Rubin, D.B (2002), *Statistical Analysis with Missing Data* (2nd edition), Wiley.

McGlincy, M.H. (2004), "A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links," *2004 Proceedings of the American Statistical Association, Survey Research Methods Section* [CD-ROM], American Statistical Association: 4001– 4008.

Runge, J.W. (2000), "Linking data for injury control research," *Annals of Emergency Medicine*, **35**, 613–615.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall.

Winkler, W. E. (1988), "Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage," *1988 Proceedings of the American Statistical Association, Survey Research Methods Section*, American Statistical Association: 667– 671.

Winkler, W. E. (1989), "Frequency-based matching in the Fellegi-Sunter model of record linkage," *1989 Proceedings of the American Statistical Association, Survey Research Methods Section*, American Statistical Association: 778–783.

Winkler, W. E. (1993), "Improved decision rules in the Fellegi-Sunter model of record linkage," *1993 Proceedings of the American Statistical Association, Survey Research Methods Section*, American Statistical Association: 274–279.

Winkler, W. E. (1994), "Advanced methods for record linkage," *1994 Proceedings of the American Statistical Association, Survey Research Methods Section*, American Statistical Association: 467–472.

Winkler, W. E. (2005), "Test Databases for Evaluating Approximate Joins," Private Communication.