# 1

# Identifying likely duplicates by record linkage in a survey of prostitutes by Thomas R. Belin, Hemant Ishwaran, Naihua Duan, Sandra H. Berry, and David E. Kanouse

## 1.1 Concern about duplicates in an anonymous survey

The Los Angeles Women's Health Risk Study (LAWHRS) was a survey of female street prostitutes in Los Angeles County that aimed to provide insight into the evolution of the AIDS epidemic in the early 1990's (Kanouse et al. 1999). Goals of the study included estimating the size of the female street prostitute population in Los Angeles, determining seroprevalence of the HIV

virus among female street prostitutes, measuring the prevalence of sexual and drug-related risk behaviors associated with HIV transmission, measuring the frequency of condom use and other preventive behaviors, and relating HIV status to behavior patterns and prostitute characteristics.

The LAWHRS was designed as a probability sample of areas of Los Angeles, times of day, and days of the week (Duan et al. 1992). The area frame was assembled from police, health officials, and study consultants including former prostitutes, and special procedures were developed for field staff (interviewers, drivers, and phlebotomists) to go through sampled areas beginning at randomly selected start points, to approach women for interviews in a systematic fashion until agreement was obtained from a woman in the area, and to obtain informed consent (Kanouse et al. 1999). Interviews typically took roughly 45 minutes, and participants were asked to provide a blood sample to be tested for exposure to HIV, syphilis, and hepatitis B. Women were paid $25 for participation. Blood-test results were not immediate, but women could obtain test results by calling RAND to arrange an appointment. Test results were stored and retrieved using a "distinguishing" code constructed at the time of interview by stringing together a set of responses to seemingly innocuous questions such as, "What is the first letter of your mother's maiden name?" We allude to the code as "distinguishing" rather than "identifying" because it succeeded in distinguishing between individuals in the study without allowing anyone to associate the information with a person in the way that a name or social security number would allow.

After eligibility for the study was established through a question about trading sex for money or drugs in the previous year, a screening question sought to avoid duplicate interviews by asking, "Have you been interviewed already by the Los Angeles Women's Health Risk Study?" An informed consent procedure was administered to respondents who acknowledged eligibility and who stated that they had not previously been interviewed. As part of the protocol, participants were told that they would not be asked to disclose their name, address, or other information that could be used to identify them personally. The informed consent form was signed by the interviewer, who certified that the respondent had reviewed all of the points on the form.

The present chapter describes an approach that was developed to address concerns that arose over possible duplicate interviews. The payment for participation was judged to be large enough that it might provide motivation for individual prostitutes to participate more than once. But because the interviews were anonymous, duplicate interviews could not be identified in a straightforward way. Some insights were possible based on the distinguishing codes associated with individual prostitutes for retrieving blood test results. Although such codes would not enable personal identification in the manner of a name or social security number, they enabled the research team to assess whether individuals had participated more than once, at least to the extent that participants would answer the questions the same way in successive in-

terviews.

The study produced 998 completed interviews, representing roughly 61% of the 1,629 women who were approached for screening. Based on the distinguishing codes, there were 55 individuals who were "known" to be interviewed more than once: 50 individuals had duplicate identifying codes in the database, and 5 individuals had codes that were observed in triplicate. However, distinguishing codes were not available for approximately 23% of the interviews (Ishwaran et al. 1991). Further, it was suspected that some subjects might have misled interviewers when answering questions to be used for identification purposes. This gave rise to concern that there were additional undetected duplicate interviews in the database, carrying the potential to bias estimates of HIV prevalence and other outcomes of interest.

The present chapter summarizes collaborative work building on methods for calibrating error rates in record linkage settings using a mixture-model framework (Belin and Rubin 1995). The effort strengthened the foundation of the LAWHRS by providing evidence that the study had indeed succeeded in interviewing a large number of different street prostitutes. In this chapter, we review relevant frameworks for record linkage that relate to the problem of identifying duplicates, after which we describe the procedure used to identify duplicates in the LAWHRS and offer comments for future applications.

## 1.2 General frameworks for record linkage

The problem of identifying duplicate records in the LAWHRS is framed here as a problem of record linkage, which generally refers to a technique for identifying individual records in one or more databases that correspond to the same person. Early theoretical work on record linkage (e.g., Newcombe et al. 1959; Fellegi and Sunter 1969) gave rise to strategies for bringing together candidate matched pairs of records. These authors recognized that there were inherent uncertainties in automated procedures requiring investigators to develop tolerances for false linkages, or false matches. Fellegi and Sunter (1969) outlined a procedure for estimating false-match rates that made use of the estimated probabilities of agreement on components within individual records that provided the basis for their procedure to assign weights characterizing closeness of agreement between record pairs. For example, typographical and transcribing errors might result in gender agreeing 99% of the time between pairs of records referring to the same person, while chance agreement would suggest that gender might agree 50% of the time between pairs of records from different people. Assuming independence of agreement across fields of information within records, one can estimate probabilities of false match as a function of the agreement weights, which can then be used to establish a cutoff between record pairs treated as matches and record pairs treated as non-matches. Belin (1993) shows that the performance of record-linkage

procedures can depend critically on decisions about where to set such a cut-off. But as noted in Belin and Rubin (1995) and Larsen and Rubin (2001), the Fellegi-Sunter approach can founder on violations of the independence assumption, giving rise to inaccurate estimates of false-match rates. (For example, in census applications considered in those articles, agreement on first name would clearly not be independent of agreement on gender.)

Belin and Rubin (1995) finesse the independence assumptions in the Fellegi-Sunter framework by tapping outside information, namely previously processed databases that have already been reviewed for accuracy by teams of matching clerks. These databases, with matching-clerk determinations taken as a gold standard, offer information regarding the distribution of agreement weights for true matched pairs and the distribution of agreement weights for false matched pairs. The key innovation of Belin and Rubin (1995) involved viewing the agreement weights in a database not yet reviewed by matching clerks as arising from a mixture of weights for true matches and weights for false matches. In this context, it is not essential that the agreement weights be derived from a procedure such as that of Fellegi and Sunter (1969); more crucial is that the procedures used to develop agreement weights are exchangeable across applications. Two-component mixture models could then be fit in a current database by making use of informative priors derived from previously processed data. Other elements of the estimation strategy include the EM algorithm (Dempster, Laird, and Rubin 1977) to obtain posterior modes for mixture-model parameters, the SEM algorithm to obtain asymptotic standard errors (Meng and Rubin 1991), and multiple imputation to average over uncertainty about appropriate normalizing transformations (Rubin 1987).

Another strategy for calibrating error rates in record linkage is described by Larsen and Rubin (2001), who extend the Fellegi-Sunter approach in a more flexible framework that allows dependence of agreement among fields of information in records. The mixture-model idea is still central, as the set of all pairs of records is partitioned based on latent indicators into separate classes, but the models in this context are mixtures for discrete data (reflecting agreement or disagreement on each of several characteristics). In the census application that served as a motivating example, three-class mixtures were explored, where the fitted models tended to divide pairs into same-household matches, same-household non-matches, and different-household non-matches. Instead of assuming the existence of a large database that has already been processed, Larsen and Rubin (2001) propose to achieve accurate calibration of false-match rates by fitting a mixture model, selecting subsets of the original record pairs to be reviewed for accuracy of matching determinations, and iterating the model-fitting and review process until only a small proportion of record pairs reviewed in successive review cycles appear to be matches.

## 1.3 Estimating probabilities of duplication in the Los Angeles Women's Health Risk Study

In the LAWHRS, the availability of 50 known duplicates and 5 known triplicates based on information in the "distinguishing codes" presented an opportunity to use the Belin and Rubin (1995) framework to assess the extent of additional duplication in the database. Specifically, the known duplicates and triplicates would provide a "training" data set where, for a given metric summarizing closeness of agreement between answers in the balance of the interview, the training data would provide information on the distribution of the agreement metric between duplicate interviews on the same person as well as information on the distribution of the agreement metric between interviews on different people. We elaborate by summarizing a procedure for choosing a distance metric, discussing the estimation of mixture components, and describing findings from the LAWHRS.

### 1.3.1 Choosing a distance metric

There were 14 questionnaire items used to construct the individual distinguishing codes in the LAWHRS. Because these items were not available on all individuals, it was necessary to use other questionnaire items to develop a distance metric to summarize closeness of agreement between records. Ishwaran et al. (1991) list 107 questionnaire items that were available for inclusion in a distance metric. While it would have been possible to use all items in the metric, it was presumed that some items would contribute substantially to the ability to distinguish records while other items would not. To avoid including items in the distance metric that were largely adding noise to the assessment, it was decided to include items in the distance metric only if there was evidence that they would contribute to the ability to distinguish individuals.

This problem was conceptualized using a testing framework to decide whether a pair of records represents two different individuals or two records from the same individual. Suppose the database has $n$ records. Let questionnaire items be indexed by $i$, where $i = 1, 2, ..., 107$, let record pairs be indexed by $j$, where $j = 1, 2, ..., n(n-1)/2$, and let $\delta_{i,j}$ represent the indicator function comparing record pair $j$ on question $i$, with the result equal to 1 if there is agreement and equal to 0 if there is disagreement. The set of record pairs can be partitioned into the set $T$ of true-matched (or duplicate) pairs and the set $F$ of false-matched pairs. If we let $p_{0i}$ represent the probability of agreement on question $i$ between two different individuals and $p_{1i}$ represent the probability of agreement on question $i$ in two interviews with the same

individual, we can write

$$\delta_{i,j} \sim \begin{cases} \text{Bernoulli}(p_{0i}) & \text{if } j \in F \\ \text{Bernoulli}(p_{1i}) & \text{if } j \in T. \end{cases}$$

The training data set provides information that can be used to estimate $p_{0i}$ and $p_{1i}$. Specifically, we can let

$$\hat{p}_{0i} = \frac{\sum_{j \epsilon F} \delta_{i,j}}{\mid F \mid}$$

and

$$\hat{p}_{1i} = \frac{\sum_{j \epsilon T} \delta_{i,j}}{\mid T \mid},$$

where $|\cdot|$ represents the cardinality of a set. We assume that the 50 record pairs with duplicate distinguishing codes and the 5 record triples with triplicate distinguishing codes in the training set refer to 55 different individuals (e.g., we assume that a person did not respond twice with one set of answers to the distinguishing-code questions and twice more with a different set of answers to the distinguishing-code questions). Then, except for the triplicates, we should have near independence among the $\delta_{i,j}$ for $j \in T$. Therefore

$$\left(|T|\,\hat{p}_{1i}\right) \sim \text{Bin}(|T|, p_{1i}),$$

approximately. However, there may be quite a bit of dependence among the $\delta_{i,j}$ when $j \epsilon F$. For instance, the first record from duplicate pair A will be compared against both records of a different duplicate pair B, and the second record from duplicate pair A will also be compared against both records from duplicate pair B. These four comparisons are apt to yield the same indicator values if the duplicates are well matched. Ignoring triplicates we should expect

$$\left(|F|\,\hat{p}_{0i}\right) \sim 4\text{Bin}(|F/4|, p_{0i}),$$

at least approximately.

For items where $p_{0i} = p_{1i}$, which would not be useful for distinguishing true-matched and false-matched pairs, we would have

$$\hat{p}_{1i} - \hat{p}_{0i} \sim \text{N}(0, s^2)$$

where

$$s = \sqrt{\frac{\hat{p}_{1i}(1-\hat{p}_{1i})}{\mid T \mid} + \frac{4\hat{p}_{0i}(1-\hat{p}_{0i})}{\mid F \mid}}.$$

based on using the fact that $\hat{p}_{1i}$ is independent of $\hat{p}_{0i}$ and applying a normal approximation.

These results were used to motivate a decision rule to include question $i$ in the metric summarizing closeness of agreement if

$$\hat{p}_{1i} > \hat{p}_{0i} + 3.1\ s.$$

which corresponds to an event in the upper 0.1 percentile of the normal reference distribution. This decision rule suggested that 55 of the original 107 items would be included in the distance metric.

The second part of the algorithm involved assigning a weight to a chosen question. If question $i$ was deemed suitable, then the weight for this question was calculated as

$$w_i = 1/2 \left( \frac{p_{1i}}{p_{0i}} + \frac{1 - p_{0i}}{1 - p_{1i}} \right).$$

The $w_i$'s ranged from values of 1.54 to 28.40. Although not formally equivalent to the weighting procedure outlined in Fellegi and Sunter (1969), which involves logarithms of the ratios of conditional probabilities of agreement given $T$ and $F$, this weighting scheme has the property of assigning high weights to questions that have a high probability of agreement under the alternative hypothesis as well as to those questions that have a high probability of disagreement under the null hypothesis. Agreement weights $Y_j$ were calculated by summing the $w_i$ values across all questionnaire items represented in record pair $j$ and rescaling so that the maximum agreement weight would equal 100.

Figure 1 displays the distribution of agreement weights for the true-matched pairs in the top display and for the false-matched pairs in the bottom display. While there is some overlap, it is clear that the metric provides a strong basis for distinguishing true-matched pairs from false-matched pairs.

In line with the findings of Belin (1993), we anticipated that even ad hoc approaches to assigning weights would perform reasonably well. For purposes of comparison, using the same questionnaire items as were included in the distance metric described above, unit weights were assigned for agreement on field $i$. As expected, this metric capturing a count of the number of fields of agreement between records produced well-separated components, but the original weighting scheme appeared to produce better separation in the region of overlap between the two components.

### 1.3.2 Estimating mixture components

Formally, the problem of determining the extent of duplication in the LAWHRS can be framed as a mixture problem, where the distribution of all distances can be partitioned into two components, one characterizing the distribution of distances associated with true-matched pairs and one characterizing the distribution of distances associated with false-matched pairs. The problem is complicated by the presence of many pairs being associated with each individual record. For example, among the 115 records in the training data

set (comprised of 50 duplicate pairs and 5 triplicate sets), one could construct 6,555 record pairs, of which 65 are true matches (the 50 duplicates plus 3 matched pairs for each triplicate), and the remaining 6,490 are false matches. The problem is further complicated by the fact that the size of the training set is small compared with the size of the set being investigated for duplicates, where there were nearly 500,000 pairs to be considered. While the distribution of weights for true matches might be fairly comparable between the training and target databases, the greater number of possible pairs in the target database raised the possibility of a different distribution of weights for false matches between the training and target databases.

A first-pass approach sought to gauge the extent of the duplication by identifying best candidate matches for each record using the newly developed closeness-of-agreement metric and then reviewing candidate matches manually. The largest weight associated with a false-matched pair in the training data set was 80.3 on the 100-point scale, so in this first pass is was decided to consider all pairs with weights above 80.3 to be duplicates. Individual distinguishing codes were also assessed to judge whether some of the cases might be part of triples or quadruples. This process yielded 25 new suspected doubles, 7 new suspected triples (4 of which were duplicates in the training set), and 2 new quadruples (1 of which was a triplicate in the training set). According to this tally, overall there were 2 quadruples, 11 triples (7 newly suspected plus 4 of the 5 triples in the training data, with the other triple in the training data now looking like a quadruple), and 71 duplicates (25 newly suspected plus 46 of the 50 duplicates in the training data, the other 4 duplicates now appearing to be parts of triples). The implied number of duplicate interviews was 3 for each of the quadruples, 2 for each of the triples, and 1 for each of the duplicates, or 99 overall. This represented roughly 10% of the 998 completed interviews.

An alternate approach to assessing the extent of contamination of the set of interviews through duplication was based on a probability model described in Belin and Rubin (1995). Working with the weights for the best candidate matches, the model assumes that the weights for true matches and weights for false matches are each normally distributed after application of two-parameter Box-Cox transformations, with distinct transformations for each component to address possibly different skewness in the weight distributions. The transformations are indexed by a power parameter $\gamma$ and a scaling parameter $g$ corresponding to the geometric mean of the observations in the following way:

$$\psi(w;\gamma,g) = \begin{cases} \frac{w^\gamma - 1}{\gamma (g)^{\gamma-1}} & \text{if } \gamma \neq 0 \\ g \log(w) & \text{if } \gamma = 0. \end{cases}$$

The two transformations are estimated from the training sample by use of a grid search of the likelihood. To facilitate identification of the mixture distribution in the target sample, the training data are also used to provide

an estimate of the ratio of the variances of the component distributions on the transformed scales. Subsequently, a mixture model is fit to the weights for best candidate matches from the target sample. An EM algorithm is available to obtain estimates of the component means and the unconstrained component variance parameter, after which posterior probabilities of duplication are available as ratios of transformed-normal component densities (Belin and Rubin 1995). Using this approach, the aggregate probability of false match (duplication) was estimated to be 14.9%. This departed somewhat from the assessment from the manual procedure, partly due to the impact of cases where the record pair was not clearly a duplicate. But the broader conclusion to investigators from both approaches was that duplication, while a concern that merited attention, was not at a level that would completely undermine findings from the very demanding fieldwork.

### 1.3.3 Results from LAWHRS

Estimates for the street-prostitute work force were obtained by combining sampling weights from the LAWHRS with estimated probabilities of duplication to downweight the impact of potential duplicates. Key findings from the LAWHRS are summarized in Berry et al. (1992) and Kanouse et al. (1992). Survey participants saw a mean of 30.2 clients per week. Vaginal sex without a condom occurred in 12% of most recent transactions, and oral sex without a condom occurred in 21% of most recent transactions, with some transactions involving both. In 30% of transactions the client requested a condom, with a condom being used in 97% of those cases. Meanwhile, in 11% of transactions the client requested that a condom not be used, with condoms not being used in roughly half of those transactions. While practical considerations involving supervision of phlebotomy delayed the initiation of blood testing, blood draws were available for over half of the sample. Laboratory analysis suggested seropositive rates of 2.5% for HIV-1 antibodies, 33% for hepatitis B surface antibodies, and 34% for past or present syphilis infection. The emerging profile suggests both that street prostitutes and their clients are at substantial risk for sexually transmitted diseases including HIV and that the amount of risk assumed is an outcome of a negotiation process.

## 1.4 Discussion

The application of Belin and Rubin's mixture-model technique for identifying duplicate interviews was successful on multiple levels in the context of this interesting applied context. First, the availability of the method provided a probabilistic framework for incorporating evidence about duplication, which was desirable in a context where great effort had been expended to obtain

a probability sample. The results, which appeared consistent with a manual approach that had face validity, helped to instill confidence not only in the method itself but, on account of the non-threatening magnitude of the estimated duplication, in the findings of the study as a whole.

Further confidence in the method derived from an anecdote that the investigators alluded to as "the search for the three-faced Eve." Feedback from field workers had identified a case as a likely triplicate, although in the clerical process, linkage was lost between the record of the third interview and two others records that could be classified as duplicates using the distinguishing codes. Although the investigators were prepared to have project staff review the hundreds of hard copies of interviews to try to find the lurking triplicate, the weighting scheme suggested a candidate match that was readily identified as the third member of the triple without requiring such extensive manual effort. This process added a measure of face validity to the methodology.

An anonymous reviewer noted the possibility of comparing characteristics of duplicates and non-duplicates to assess potential systematic relationships. That is, to build on terminology from Rubin (1976), one could consider whether records were "duplicated completely at random," with no systematic differences between duplicates and non-duplicates, or were "duplicated at random," allowing the possibility that duplicates and non-duplicates may differ on covariates. One could further consider defining weighting classes based on covariate data and using weighting adjustments to assess whether certain quantities of interest, such as AIDS prevalence, are disproportionately affected by duplication.

A final comment concerns implications of this methodology in disclosure avoidance problems. Survey organizations routinely offer pledges of confidentiality to survey participants, yet there is often considerable interest in having data files from censuses and annual surveys available for public use. The concern raised by the record-linkage methods described here is that another avenue might become available for identifying individuals in public-use data files by aggregating information on seemingly innocuous characteristics. Many individual in the United States would be uniquely identifiable given a set of, say, 50 pieces of covariate information. As a hypothetical, suppose that a data user knew 50 pieces of covariate information on an individual from public sources (e.g., white, male, age 55, city of residence, etc.) and suppose that a public use data file supplies records of personal income along with all of the same 50 items except city of residence. In such a setting, record-linkage techniques might be used by adversarial individuals to try to break confidentiality in public-use files. The implied challenge may require not just that imputation be used for disclosure avoidance, as suggested in Rubin (1993) and Raghunathan, Reiter, and Rubin (2003), but also that the imputation procedures scramble covariate information across individuals rather than just drawing entire individual records using hot-deck or approximate Bayesian bootstrap procedures. Raghunathan, Reiter, and Rubin (2003) recognize that

choices between model-based and resampling-based imputation procedures involve tradeoffs affecting both precision and protection against disclosure. Guarding against record-linkage technology implies another layer of challenge in disclosure-avoidance problems. The idea of joining seemingly distinct statistical frameworks into a unified whole, which paid off in assessing the extent of duplication in the Los Angeles prostitute survey, might also be important to a successful disclosure-avoidance strategy.

# Bibliography


[1] Belin, T.R. (1993). Evaluation of sources of variation in record linkage through a factorial experiment. *Survey Methodology*, **19**, 13–29.

[2] Belin, T.R., Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, **90**, 694–707.

[3] Berry, S.H., Kanouse, D.E., Duan, N., Lillard, L.A. (1992). Risky and non-risky sexual transactions with clients in a Los Angeles probability sample of female street prostitutes. *VIII International Conference on AIDS/III STD World Congress, Poster Abstracts, Vol. 2*, PoD 5604, Amsterdam.

[4] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **39**, 1–38.

[5] Duan, N., Kanouse, D.E., Berry, S.H. (1992). Weighting a probability sample of street prostitutes. Unpublished technical report, RAND Corporation.

[6] Fellegi, I.P., Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**, 1883–1210.

[7] Ishwaran H., Berry, S., Duan, N., Kanouse, D. (1991). Replicate interviews in the Los Angeles Women's Health Risk Study: searching for the three-faced Eve. Unpublished technical report, RAND Corporation.

[8] Kanouse, D.E., Berry, S.H., Duan, N., Lever, J., Carson, S., Perlman, J.F., Levitan, B. (1999). Drawing a probability sample of female street prostitutes in Los Angeles County. *Journal of Sex Research*, **36**, 45–51.

[9] Kanouse, D.E., Berry, S.H., Duan, N., Richwald, G., Yano, E.M. (1992). Markers for HIV-1, hepatitis B, and syphilis in a probability sample of street prostitutes in Los Angeles County, California. *VIII International Conference on AIDS/III STD World Congress, Poster Abstracts, Vol. 2*, PoC 4192, Amsterdam.

[10] Larsen, M.D., Rubin, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, **96**, 32–41.

[11] Meng, X.L., Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, **86**, 899–909.

[12] Newcombe, H.B., Kennedy, J.M., Axford, S.J., James, A.P. (1959). Automatic linkage of vital records. em Science, **130**, 954–959.

[13] Raghunathan, T.E., Reiter, J.P., Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, **19**, 1–16.

[14] Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **3**,581–592.

[15] Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley.

[16] Rubin, D.B. (1993). Satisfying confidentiality constraints through use of synthetic multiply-imputed microdata. *Journal of Official Statistics*, **9**, 461–468.