The Annals of Applied Statistics
2019, Vol. 13, No. 3, 1753–1790
https://doi.org/10.1214/19-AOAS1256

INCORPORATING CONDITIONAL DEPENDENCE IN LATENT CLASS MODELS FOR PROBABILISTIC RECORD LINKAGE: DOES IT MATTER?[1]

BY HUIPING XU[*], XIAOCHUN LI[*], CHANGYU SHEN[†], SIU L. HUI[‡] AND SHAUN GRANNIS[*,‡]

Indiana University[*], Harvard Medical School[†] and Regenstrief Institute[‡]

The conditional independence assumption of the Felligi and Sunter (FS) model in probabilistic record linkage is often violated when matching real-world data. Ignoring conditional dependence has been shown to seriously bias parameter estimates. However, in record linkage, the ultimate goal is to inform the match status of record pairs and therefore, record linkage algorithms should be evaluated in terms of matching accuracy. In the literature, more flexible models have been proposed to relax the conditional independence assumption, but few studies have assessed whether such accommodations improve matching accuracy. In this paper, we show that incorporating the conditional dependence appropriately yields comparable or improved matching accuracy than the FS model using three real-world data linkage examples. Through a simulation study, we further investigate when conditional dependence models provide improved matching accuracy. Our study shows that the FS model is generally robust to the conditional independence assumption and provides comparable matching accuracy as the more complex conditional dependence models. However, when the match prevalence approaches 0% or 100% and conditional dependence exists in the dominating class, it is necessary to address conditional dependence as the FS model produces suboptimal matching accuracy. The need to address conditional dependence becomes less important when highly discriminating fields are used. Our simulation study also shows that conditional dependence models with misspecified dependence structure could produce less accurate record matching than the FS model and therefore we caution against the blind use of conditional dependence models.

1. Introduction.

Record linkage identifies and matches records belonging to the same entity from disparate data sources (Christen (2012)). Such a task is not trivial when there is a lack of a unique identifier across data sources. Additional

Received May 2016; revised January 2019.

[1] Supported by grants R01HS018553 and R01HS023808 from the Agency for Healthcare Research and Quality and ME-2017C1-6425 from the Patient-Centered Outcomes Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality or the Patient-Centered Outcomes Research Institute.

Key words and phrases. Conditional dependence, finite mixture, Gaussian random effects model, latent class analysis, log-linear model, record linkage.

information common in multiple data sources are needed to link records together (Sadinle (2017)). Fields describing the identifying details of these records, such as their name, address, phone number, and so on, are used in record linkage. Record pairs are formed by comparing values of the fields of two records and the comparison results are used to determine whether the two records belong to the same entity. If they do, the record pair is called a match; otherwise, it is called a nonmatch. Most statistical techniques used for record linkage are based on probabilistic methods first proposed by Newcombe and Kennedy (1962) and later formalized by Fellegi and Sunter (1969). Alternative approaches include deterministic methods (Gomatam et al. (2002), Tromp et al. (2011)), Bayesian methods (Fortini et al. (2001, 2002), Larsen and Rubin (2001), Larsen (2004, 2012), Tancredi and Liseo (2011), Sadinle (2014, 2017)), unsupervised and supervised machine learning methods such as clustering, decision trees, and support vector machines (Bilenko and Mooney (2003b, 2003a), Abril, Navarro-Arribas and Torra (2012), Christen (2008), Christen (2012), Han et al. (2004), Martins (2011), Torra, Navarro-Arribas and Abril (2010), Treeratpituk and Giles (2009), Ventura and Nugent (2014), Ventura, Nugent and Fuchs (2015)), and methods that link more than two databases (Sadinle and Fienberg (2013)).

1.1. *Statistical challenges in probabilistic record linkage.* As an unsupervised classification algorithm, the Fellegi and Sunter (FS) model has demonstrated reasonable performance without the need for a training set and therefore has been a core component of probabilistic linkage algorithms and has been widely used. The FS model makes the conditional independence assumption: the individual fields' agreement patterns are independent given the true but unknown match status. This is a rather strong assumption, which may not always hold in practice (Winkler (1989), Thibaudeau (1993)). For example, if two records agree on telephone number, then they are more likely to also agree on fields such as street name and zip code, regardless of whether they are a match. Efforts addressing conditional dependence have focused on the use of the log-linear latent class model due to its flexibility to incorporate conditional dependence by adding interaction terms to the FS model in the log-linear representation (Armstrong and Mayda (1992), Larsen (1997), Larsen and Rubin (2001), Thibaudeau (1993), Tromp et al. (2008), Winkler (1993), Zhu et al. (2010), Daggy et al. (2013)). Other approaches addressing the field dependence include the Gaussian random effects latent class model (Daggy et al. (2014)).

Despite efforts to account for conditional dependence, few studies have examined whether incorporating conditional dependence improves the record matching accuracy. One study considered a situation with a small number of record pairs and only three fields (Kelly (1986)); another study looked into the problem at an extremely low match prevalence (Tromp et al. (2008)). A recent study performed a more thorough investigation via simulation, with true match prevalence ranging from low to high and strength of conditional dependence varying from weak to strong (Daggy et al. (2014)). Findings of these studies provided preliminary evidence that incorporating conditional dependence might improve matching accuracy. However, it is not clear whether and how such improvement is affected by factors such as the match prevalence, discriminating power of fields, sample size, and class-specific conditional dependence structures.

1.2. *Conditional dependence in latent class analysis.* The FS model belongs to the family of statistical methodology known as latent class models (Li and Shen (2013)), which have been widely used. One application area is the evaluation of diagnostic test accuracy in the absence of a gold standard test, a diagnostic test that gives 100% accurate results (Walter et al. (2012)). In these problems, one wishes to evaluate a new diagnostic or screening test when there is no perfect gold standard for comparison. The robustness of the latent class model that assumes independence of diagnostic test results conditional on the true disease status has been extensively studied. Findings have shown that the conditional independence model can lead to seriously biased parameter estimates when the conditional independence assumption is not valid (Vacek (1985), Torrance-Rynard and Walter (1997)). Researchers have proposed many approaches to address conditional dependence. These include what have been used in the record linkage literature, such as log-linear latent class models (Espeland and Handelman (1989), Hagenaars (1988), Yang and Becker (1997), Xu, Black and Craig (2013)) and Gaussian random effects latent class models (Hadgu and Qu (1998), Qu, Tan and Kutner (1996), Uebersax (1999), Dendukuri and Joseph (2001)), as well as those that have not yet been used in record linkage, such as the finite-mixture extended latent class models (Albert and Dodd (2004), Albert, McShane and Shih (2001), Albert (2009)) and probit latent class models (Xu and Craig (2009), Xu, Black and Craig (2013)); See Collins and Huynh (2014) for a thorough review of latent class models from both the frequentist and Bayesian perspectives on the estimation of diagnostic test accuracy.

In the evaluation of diagnostic test accuracy, the focus of the latent class analysis is to estimate the accuracy of each diagnostic test. Record linkage problems, on the other hand, focus on the accurate prediction of the unknown match status of record pairs. Record linkage methods should therefore be evaluated and compared in terms of record matching accuracy. Despite the significant difference between these two application areas, methodological findings obtained from diagnostic test evaluations may still have critical implications in record linkage. When applied to record linkage problems, parameters in latent class models represent the match prevalence and the m- and u-probabilities (the m-probabilities are the probabilities of the field agreement given that the record pair belongs to the same entity; the u-probabilities are the probabilities of field agreement given that the two records do not belong to the same entity). These probabilities are at the heart of the probabilistic record linkage: the matching weights are defined as the logarithms of the m- and u-probability ratios, and the sum of the weights across fields, known as the matching score, is used to declare matches and nonmatches. The m- and u-probability ratio of a field quantifies the "discriminating power" of the field. The larger the ratio is, the bigger contribution the field makes to the matching score. Biases in these parameter estimates could therefore have a large impact on record matching accuracy.

1.3. *Overview of the paper.* The purpose of the paper is to investigate whether and how the matching accuracy is improved by accommodating the conditional dependence among matching fields in latent class models compared to the conditional independence FS model. The remainder of the paper is organized as follows. Section 2 introduces a motivating example: linking patient records from two hospitals. In Section 3, we present the existing record linkage approaches and extend the existing models using the finite mixture idea from the diagnostic testing literature. Model estimation is implemented in the standard software SAS with the NLMIXED procedure described in Section 3.3. These models are then applied to the hospital data linkage example and their matching accuracies are compared against the manual review results in Section 4. Two additional examples involving deduplication of cancer registry data and disambiguation of inventor records are used to further evaluate the influence of accounting for conditional dependence, where results are presented in Appendix B and Appendix C. A simulation study is conducted in Section 5 to investigate the impact of accommodating conditional dependence on matching performance and how it may be affected by factors including the match prevalence, discriminating power of fields, sample size, and class-specific conditional dependence structures. These are followed by concluding remarks in Section 6.

2. A motivating example.

To motivate our study, we consider a real-world use case that links records from patient registries of two hospitals in central Indiana. One hospital is a public inner city hospital system with a large underserved population and the other is a private urban hospital system serving a larger population. Because of the close physical proximity of the two hospitals, many patients receive care and many physicians provide care across both hospitals. Thus, clinical data for these care processes are fragmented across the systems. The purpose of linking patient records from these two hospitals is to ensure that patients' clinical data are complete to best inform medical decision making and care coordination. These two hospitals contain approximately 1.5 million and 3.8 million patient records, respectively. Fields used for identifying matches in these two sets of data include first name, last name, middle initial, day, month, and year of birth, sex, zip code, phone number, and social security number (SSN). Fields such as first and last names, sex, and birth date are relatively complete, while other fields contain more missing data. For example, the SSN is missing for approximately one third of the hospital records, hence producing a large number of record pairs with missingness on the agreement of the SSN field. In record linkage, missing data have been addressed by imputation (Herzog, Scheuren and Winkler (2007)), adding a third category in addition to the binary agreement/disagreement pattern in the latent class modeling (Tromp et al. (2008)), or treating fields with missingness as a disagreement (Ong et al. (2014)). We use the third approach since it is more conservative and protects against false matches.

Due to the extremely large number of record pairs, record linkage analysis usually employs a blocking scheme to decrease the number of record pairs in consideration (Herzog, Scheuren and Winkler (2007)). One or more fields are taken as blocking variables and records with agreement on blocking variables are compared and evaluated for their true match status. Records with disagreement on blocking variables are automatically considered as nonmatches. Since disagreement in blocking variables may arise from erroneous values of the variables, record linkage practitioners typically use multiple blocking schemes, starting with more restrictive blocking variables such as the SSN, and using less restrictive blocking variables subsequently. In our example, we used two blocking schemes: one with the SSN as the blocking variable, representing a rather restrictive scheme, and the other with last name and first name (LNFN) as blocking variables, representing a less restrictive scheme. The SSN blocking scheme produced a data set of 590,128 record pairs (the SSN block) and the LNFN blocking scheme produced a data set of 14,665,148 record pairs (the LNFN block). After blocking, a random sample of record pairs was selected from all record pairs resulted from each blocking scheme and manually reviewed by five reviewers to establish the true match status. Both samples consisted of approximately 5500 record pairs and were selected via stratified sampling that over-sampled record pairs with more ambiguity on the matching status, reflected by greater inconsistency in the agreement status of matching fields. Once the sample of record pairs was selected, manual review was implemented using a balanced incomplete block design. Each reviewer reviewed 40% of the record pairs so that every record pair was reviewed by two reviewers. Discrepancies were adjudicated by a third reviewer. Approximately 7.6% of the manually reviewed record pairs in the SSN blocking scheme and 4% in the LNFN blocking scheme yielded inconsistent results from the two reviewers and hence required a third adjudicator. As expected, inconsistent reviews occurred more frequently for pairs with greater discrepancy in the agreement status of fields.

These two blocking schemes are chosen to be presented because they represent situations with very low and very high match prevalence. Based on the manual review sample, 97% of the record pairs in the SSN blocking scheme are true matches, but only 6% of the record pairs in the LNFN blocking scheme are. In the SSN blocking scheme, we consider fields with relatively poor discriminating power and high conditional dependence, with the purpose to evaluate whether incorporating conditional dependence helps to improve the matching accuracy when matching fields have poor discrimination. We therefore choose fields including zip code, telephone number, middle initial, last name and sex. In the LNFN blocking scheme, we consider fields with good discriminating power including SSN, telephone number, zip code, day, month and year of birth. These two blocking schemes allow us to examine whether incorporating conditional dependence matters when there is a dominating class and whether the discriminating power of matching fields plays a role.

Binary agreement/disagreement pattern is considered for the comparison of each field since the FS model is based on binary data (Fellegi and Sunter (1969)). For first and last names, the modified Jaro–Winkler string comparator was used to obtain the similarity measure, which was then dichotomized using a threshold of 0.8. This threshold was selected based on our empirical experience and previous research, which showed that, at a threshold of 0.8, the modified Jaro–Winkler comparator achieved highest linkage sensitivity of 97% (Grannis, Overhage and McDonald (2004)). With dichotomous data on the agreement/disagreement of each field, there are $2^5 = 32$ unique vector patterns in the SSN blocking scheme (see Table 2 in Appendix A) and $2^6 = 64$ unique vector patterns in the LNFN blocking scheme (See Table 3 in Appendix A). Based on the manual review sample for each blocking scheme, we estimate the true match prevalence and m- and u-probabilities (Table 1). The conditional independence of the data is examined using the correlation residual plot, shown in Figure 1, where the correlation residual is defined as the difference in the observed correlation and expected correlation for each pair of fields (Qu, Tan and Kutner (1996)). The observed correlation is calculated based on the observed frequencies of record pairs in the 2×2 contingency table formed by a pair of fields. Following the recommendation by Subtil, de Oliveira and Gonçalves (2012), the expected correlation is calculated based on the expected frequencies estimated in the FS modeling framework, using the parameters established by the manual review. If fields are conditionally independent, the differences in the observed and expected correlations will be nearly zero. On the other hand, if two fields are conditionally dependent, the observed and expected correlations will be different, resulting in a large deviation from zero on the correlation residual plot.

For both SSN and LNFN blocking schemes, the conditional independence assumption is clearly invalid and the FS model is inadequate. This can be seen from the considerable differences between the observed and the expected correlations shown in the correlation residual plots. The largest correlation residual is between telephone number and zip code for both blocking schemes. These two fields are conditionally dependent potentially because subjects with the same phone number are more likely to have the same zip code since they are likely the same person, or different persons residing in the same household. In the SSN blocking scheme, correlation residuals between last name and zip code and between last name and telephone number are also relatively large, indicating conditional dependence among

TABLE 1
Estimates of match prevalence and m- and u-probabilities of latent class models for the hospital linkage data

	Manual review	FS	FSFM	LL	LLFM	GRE	GREFM
The SSN blocking scheme							
Prevalence	0.9695	0.4415	0.4364	0.9548	0.7151	0.9258	0.7768
m-probabilities							
Zip code	0.4718	0.8477	0.8474	0.4686	0.5192	0.4776	0.5011
Telephone number	0.2551	0.5282	0.5338	0.2471	0.2890	0.2528	0.2758
Middle initial	0.1658	0.2057	0.2010	0.1729	0.2172	0.1782	0.2019
Last name	0.8891	0.9462	0.9443	0.8806	0.9443	0.8982	0.9224
Sex	0.9940	0.9941	0.9921	0.9958	0.9962	0.9953	0.9965
u-probabilities							
Zip code	0.1877	0.1450	0.1516	0.1729	0.2945	0.1764	0.2956
Telephone number	0.0347	0.0095	0.0099	0.0582	0.1118	0.0601	0.1086
Middle initial	0.0214	0.1346	0.1389	0.0203	0.0375	0.0142	0.0411
Last name	0.2534	0.7874	0.7903	0.3696	0.6395	0.3497	0.6316
Sex	0.5776	0.9701	0.9718	0.6608	0.9418	0.7977	0.9257
Deviance		10,706.5	1782.9	163.8	29.92	447.6	47.4
The LNFN blocking scheme							
Prevalence	0.0620	0.0598	0.0597	0.0601	0.0596	0.0601	0.0596
m-probabilities							
Birth year	0.9877	0.9779	0.9781	0.9761	0.9796	0.9761	0.9796
SSN	0.5070	0.5290	0.5291	0.5257	0.5295	0.5254	0.5293
Birth day	0.9556	0.9756	0.9756	0.9747	0.9779	0.9747	0.9779
Telephone number	0.1921	0.2010	0.2007	0.1987	0.2002	0.1984	0.1999
Zip code	0.3626	0.3978	0.3972	0.3936	0.3958	0.3931	0.3953
Birth month	0.9650	0.9891	0.9891	0.9893	0.9909	0.9893	0.9909
u-probabilities							
Birth year	0.0203	0.0185	0.0185	0.0183	0.0186	0.0183	0.0186
SSN	0.0019	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
Birth day	0.0333	0.0322	0.0322	0.0319	0.0322	0.0319	0.0322
Telephone number	0.0000	0.0004	0.0005	0.0005	0.0005	0.0005	0.0005
Zip code	0.0093	0.0117	0.0117	0.0118	0.0119	0.0118	0.0119
Birth month	0.0755	0.0816	0.0816	0.0813	0.0817	0.0813	0.0817
Deviance		297,041.4	147,527.4	19,866.76	15,056.4	18,534.8	13,770.56

these fields. Subjects with the same last name (and SSN) are likely the same person or different persons in the same household (family members may share SSN), and hence are more likely to have the same telephone number or zip code. This is likely one of the reasons for the conditional dependence between SSN and zip code and between SSN and telephone number in the LNFN blocking scheme.

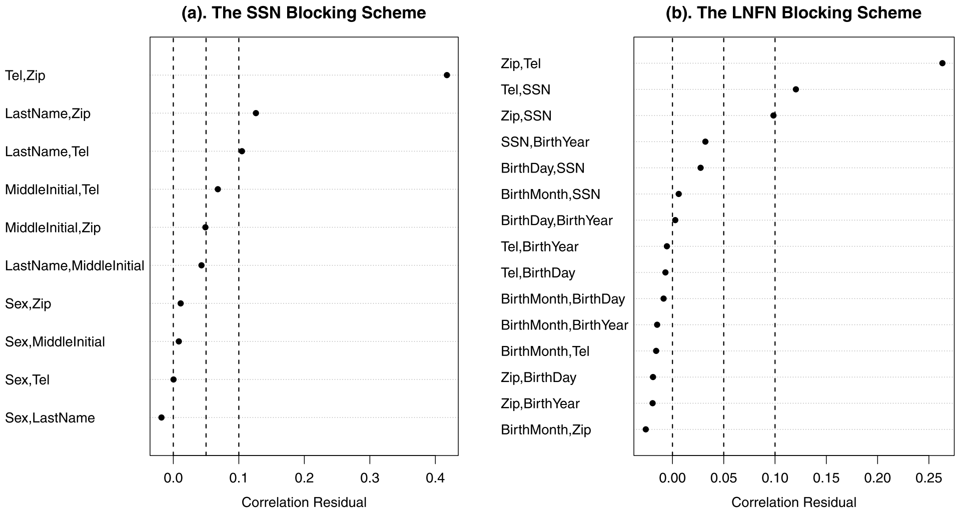

FIG. 1. *Correlation residual plot of the hospital linkage data for* (*a*) *the SSN blocking scheme* (*fields include zip code* (*Zip*), *telephone number* (*Tel*), *middle initial* (*MiddleInitial*), *last name* (*LastName*), *and sex*) *and* (*b*) *the LNFN blocking scheme* (*fields include year of birth* (*BirthYear*), *social security number* (*SSN*), *day of birth* (*BirthDay*), *telephone number* (*Tel*), *zip code* (*Zip*), *and month of birth* (*BirthMonth*)). *For each pair of fields*, *the plot shows the difference between observed correlation and expected correlation*, *calculated based on the observed frequencies and expected frequencies estimated by the FS model. Nonzero correlation residuals indicate the inadequacy of the FS model.*

3. Probabilistic record linkage.

3.1. *Existing approaches*. Let $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_J)$ be the agreement vector of the J fields for a record pair, where $Y_j = 1$ if the two records agree on the jth field and 0 otherwise. Let M denote the true match status of the record pair, which takes values 0 (true nonmatch) and 1 (true match). Based on the observed agreement pattern $\boldsymbol{Y}$, the contribution of the record pair to the likelihood is

$$P(\boldsymbol{Y}) = P(\boldsymbol{Y}|M = 1)P(M = 1) + P(\boldsymbol{Y}|M = 0)P(M = 0), \tag{1}$$

where $P(M = 1)$, denoted as π, is the prevalence of the true matches and $P(\boldsymbol{Y}|M)$ is the conditional distribution of the agreement vector $\boldsymbol{Y}$ given the true match status M.

There are various ways to characterize the conditional distribution $P(\boldsymbol{Y}|M)$. The FS model of record linkage assumes that the agreement patterns on multiple fields of the same record pair are independent conditional on the true match status. In other words, $P(\boldsymbol{Y}|M) = \prod_{j=1}^{J} P(Y_j|M)$, where the conditional probabilities $P(Y_j|M)$ define the m- and u-probabilities $m_j = P(Y_j = 1|M = 1)$ and $u_j = P(Y_j = 1|M = 0)$.

Due to the potential violation of the conditional independence assumption in real-world record linkage applications, the FS model typically provides an inadequate fit to the data. A frequently used approach to relax the conditional independence assumption is the log-linear (LL) latent class model. It has been shown that the FS model can be equivalently parameterized using the log-linear modeling framework as follows (Clogg (1995)):

$$\log\bigl(P(\boldsymbol{Y}|M)\bigr) = \lambda + \lambda_M I(M=1) + \sum_{j=1}^{J} \lambda_{j1} I(Y_j = 1) I(M=1) + \sum_{j=1}^{J} \lambda_{j0} I(Y_j = 1) I(M=0),$$

where $I(\cdot)$ is an indicator function. Using this formulation, we can easily extend the conditional independence by including interaction terms between agreement patterns in two or more fields (with or without interacting with M). For example, if fields 2 and 3 are conditionally dependent in the true match class, we can include their interaction term in the model as follows:

$$\log\bigl(P(\boldsymbol{Y}|M)\bigr) = \lambda + \lambda_M I(M=1) + \sum_{j=1}^{J} \lambda_{j1} I(Y_j = 1) I(M=1) + \sum_{j=1}^{J} \lambda_{j0} I(Y_j = 1) I(M=0) + \lambda_1^{23} I(Y_1 = 1) I(Y_2 = 1) I(M=1).$$

More recently, the Gaussian random effects (GRE) model was applied to the record linkage literature to accommodate the conditional dependence (Daggy et al. (2014)). They introduced a random effect T that followed a standard normal distribution and assumed that the binary agreement pattern Y_j for an individual field was independent Bernoulli with proportion $\Phi(a_{jM} + b_{jM}T)$, conditional on the true match status M and random effect T. Under the GRE model, the m- and u-probabilities can be calculated as follows:

$$m_j = \Phi\biggl(\frac{a_{j1}}{\sqrt{1 + b_{j1}^2}}\biggr), \qquad u_j = \Phi\biggl(\frac{a_{j0}}{\sqrt{1 + b_{j0}^2}}\biggr).$$

3.2. *Latent class models with a finite mixture* (*FM*) *extension.* The GRE model incorporates possible conditional dependence among fields using a continuous mixture model framework. It uses a normal random effect to represent the heterogeneity across record pairs in the same match class. In the diagnostic testing literature, Albert and colleagues proposed an alternative approach to incorporating conditional dependence with a finite mixture model framework to handle situations where some truly diseased and nondiseased individuals were always diagnosed correctly and others were subject to diagnostic error (Albert, McShane and Shih (2001), Albert and Dodd (2004, 2008)). This model has been used to extend both the conditional independence models and the GRE models. It has been shown to produce better goodness-of-fit to the data than the conditional independence models and provide less biased parameter estimates when the conditional independence model did not provide a satisfactory fit.

In record linkage, latent class models can be similarly extended using the finite mixture model framework. We assume that record pairs in the match class contain two subclasses, one with high data quality whose field agreements are consistent with the underlying truth, and the other with relatively lower data quality whose field agreements are subject to error. Likewise, the nonmatch class also contains two subclasses of record pairs, one with field agreements of record pairs consistent with the unobserved truth and the other with field agreements subject to error. Specifically, for each record pair, let L be an indicator of whether field agreement pattern is *always* consistent with its latent truth. For a record pair, if its agreement vector $\boldsymbol{Y}$ contains both 1's and 0's (some fields agree while others do not), then $L = 0$. On the other hand, if the vector $\boldsymbol{Y}$ contains all 1's or all 0's, the record pair could potentially belong to the subclass whose vector pattern is always consistent with the true match status. Altogether, the population of record pairs is hypothesized to contain four potential classes based on M (0 or 1) and L (0 or 1). The probability of the observed agreement pattern $\boldsymbol{Y}$ is given by

$$\begin{aligned} P(\boldsymbol{Y}) = {} & \pi\eta_1 P(\boldsymbol{Y}|M=1, L=1) + \pi(1-\eta_1)P(\boldsymbol{Y}|M=1, L=0) \\ & + (1-\pi)\eta_0 P(\boldsymbol{Y}|M=0, L=1) \\ & + (1-\pi)(1-\eta_0)P(\boldsymbol{Y}|M=0, L=0), \end{aligned}$$

where $\eta_1 = P(L=1|M=1)$ and $\eta_0 = P(L=1|M=0)$ are the proportions of true matches and true nonmatches whose field agreements are always consistent with the truth. The conditional probability of $\boldsymbol{Y}$ given M and L is:

$$P(\boldsymbol{Y}|M=1, L=1) = \begin{cases} 1 & \text{if } Y_1 = Y_2 = \cdots = Y_J = 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$P(\boldsymbol{Y}|M=0, L=1) = \begin{cases} 1 & \text{if } Y_1 = Y_2 = \cdots = Y_J = 0, \\ 0 & \text{otherwise,} \end{cases}$$

and the probabilities $P(\boldsymbol{Y}|M, L=0)$ can be modeled using any approaches including the FS, LL and GRE models. For example, the m- and u-probabilities of the finite mixture extended FS model, denoted as the FSFM model, can be computed as follows:

$$\begin{aligned} m_j &= \eta_1 + (1-\eta_1)P(Y_j = 1|M=1, L=0), \\ u_j &= (1-\eta_0)P(Y_j = 1|M=0, L=0). \end{aligned}$$

These probabilities can be similarly calculated for the finite mixture extended LL and GRE models, denoted as the LLFM and GREFM models, respectively.

3.3. *Model estimation.* With J fields, there are $K = 2^J$ possible unique vector patterns. The log-likelihood of the data is given by

$$l = \sum_{k=1}^{K} f_k \ln\{P(\boldsymbol{Y}_k)\}, \tag{2}$$

where f_k is the frequency of record pairs with vector pattern $\boldsymbol{Y}_k$ and $P(\boldsymbol{Y}_k)$ is the probability of the vector pattern in (1).

The maximum likelihood estimates (MLE) of parameters are obtained by maximizing the log-likelihood function (2). In record linkage, the expectation maximization (EM) algorithm is commonly used to obtain the MLE of the model parameters (Winkler (1988)). Recently, the Newton Raphson type routines implemented in the SAS NLMIXED procedure have been used to find the MLE in record linkage (Daggy et al. (2013)). We adopt this approach for parameter estimation of the finite mixture extension of the latent class models. For the GRE model and its finite mixture extension, we will use the adaptive Gaussian quadrature approach to estimate the log-likelihood. Estimation of the matching score requires some additional programming using the NLMIXED procedure. SAS codes for fitting the latent class models for the real-world linkage example are available at http://pages.iu.edu/~huipxu/publications.html.

3.4. *Classification of record pairs.* Once parameter estimates of the latent class models are obtained, the match score, defined as $\log_2\{P(\boldsymbol{Y}|M = 1)/P(\boldsymbol{Y}|M = 0)\}$, can be estimated for each record pair and used for the classification of record pairs. High match scores indicate a greater likelihood to be a match and low match scores indicate a greater likelihood to be a nonmatch. In practice, record linkage practitioners typically use a two-threshold scheme. Record pairs with a match score above the upper threshold are classified as matches and those with a match score below the lower threshold are classified as nonmatches (Fellegi and Sunter (1969)). Record pairs with match scores falling in between the two thresholds require human review to evaluate their match status. In situations where human review is not possible due to the limited resources or privacy concern, a one-threshold scheme has also been used to classify record pairs as matches if their match scores are above the threshold and as nonmatches otherwise (Grannis et al. (2003)). For ease of comparison, we will evaluate the models using the one-threshold scheme in our paper.

4. Hospital data linkage. In this section, we will apply latent class models to the hospital data linkage example described in Section 2 and compare their performance based on the true match status established through manual review. We first fit the FS model to the data, ignoring its invalid conditional independence assumption. Further examination of the correlation residual plot in Figure 1 shows that the correlation residuals are relatively large among the following fields in the SSN blocking scheme: telephone number, zip code, last name and the middle initial. In the LNFN blocking scheme, the correlation residuals are well above the 5% level among three fields: telephone number, zip code and the SSN. The conditional dependence will be addressed using both the LL model with pairwise interactions and the GRE model with nonzero b_{jM} among these fields. Finite mixture extensions of the FS, LL and GRE models are also fit to accommodate the conditional dependence among fields. Parameter estimates of the match prevalence, m- and u-probabilities are shown in Table 1, with goodness of fit of the models presented using the deviance. In addition, we compute the marginal probabilities $P(\boldsymbol{Y})$ for each vector pattern and present the expected frequencies estimated from each model along with the observed frequencies in Table 2 and Table 3 in Appendix A. Large differences between the observed and expected frequencies indicate the lack of fit of models.

To evaluate the matching accuracy of the latent class models, we construct the Receiver Operating Characteristic (ROC) curve and compute the area under the curve (AUC) based on the true match status established in the manual review sample. Specifically, we first calculate the match score for each record pair in the manual review sample. We then compute the true positive rate and false positive rate for each threshold of the match score by comparing the model estimated match status with the true match status. The true positive rate is the proportion of truly matched record pairs that are correctly classified as matches and the false positive rate is the proportion of truly nonmatched record pairs that are incorrectly classified as matches. The ROC curves are shown in Figure 2.

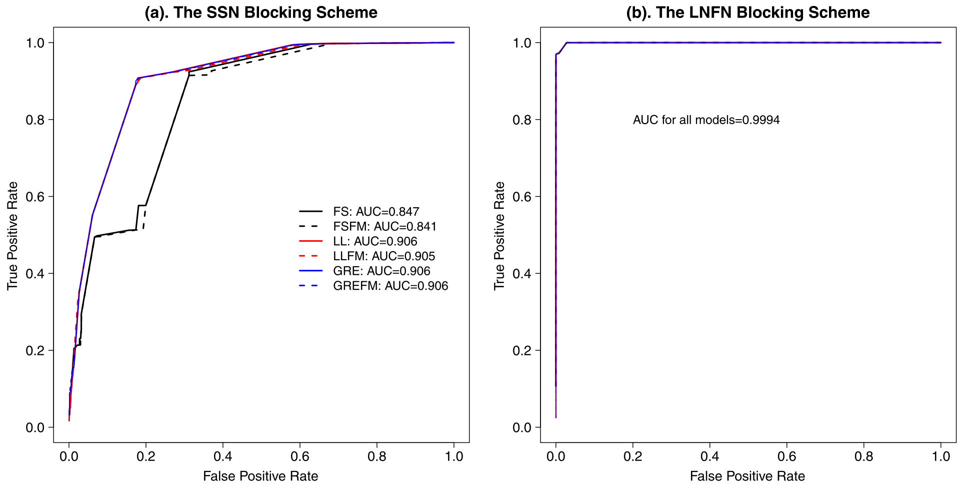

FIG. 2. *ROC curves of latent class models of the hospital linkage data for* (*a*) *the SSN blocking scheme and* (*b*) *the LNFN blocking scheme.*

4.1. *The SSN blocking scheme.* Applying the FS model effectively disregards the conditional dependence of fields. This results in a substantial underestimation of the match prevalence. According to Table 1, the SSN blocking scheme contains approximately 97% true matches, while the match prevalence is only estimated to be 44% by the FS model. Ignoring the conditional dependence among fields also leads to severely overestimated m-probabilities, especially for fields that are highly conditionally dependent. For example, the m-probability for zip code is 0.47 based on the manual review sample, implying that 47% of the truly matched record pairs agree on zip code. However, the estimated m-probability for zip code has more than doubled the true value. On the other hand, the u-probability for last name is 0.25, implying that 25% of the truly nonmatched record pairs agree on last name. This u-probability estimate has more than tripled based on the FS model. The lack of fit of the FS model can also be seen from its large deviance of 10,706.5 and the substantial differences between its expected frequencies and the observed frequencies. For example, the vector pattern with disagreement in all five fields contains more than 5000 record pairs, while the FS model estimates only approximately 1500 record pairs for this vector pattern (see Table 2 in Appendix A).

Next, we incorporate the conditional dependence using a finite mixture extended FS model. The FSFM model provides a much better fit to the data with a substantially smaller deviance of 1782.9. The expected frequencies are closer to the observed frequencies than the FS model. Despite the much improved model fit, however, there is little improvement in the estimates of match prevalence and m- and u-probabilities. Using the LL and GRE models, the model fit is further improved with deviances of 163.8 and 447.6, respectively. The expected and observed frequencies are similar for all vector patterns. More importantly, the estimated match prevalence and m- and u-probabilities are very close to their true values estimated based on the manual review. The match prevalence is estimated to be above 90% for both LL and GRE models, less than 5% away from the true value. As the LL and GRE models are further extended using a finite mixture model framework, the model fit continues to improve with deviance lower than 50 and the differences between the expected and observed frequencies become smaller. Surprisingly, the estimated match prevalence, the m- and u-probabilities move away from the true values despite the better fit of the two models. An explanation could be that the conditional dependence structures imposed by the finite mixture extended models may not be correct.

With regard to the matching accuracy, the Receiver Operating Characteristic (ROC) curves in Figure 2 show that the four models, the LL and GRE models and their finite mixture extensions, produce similar matching accuracies with AUCs of approximately 90%. The matching accuracies of the FS and FSFM models are lower with AUCs of approximately 85%. These results show that for the SSN blocking scheme, incorporating the conditional dependence improves the model fit dramatically. It also leads to improved parameter estimation and better matching accuracy.

4.2. *The LNFN blocking scheme*. The estimated match prevalence and m- and u-probabilities based on the FS model are very similar to the true values based on the manual review, despite the tremendous lack of fit indicated by the large deviance of 297,041.4 (Table 1). Conditional dependence models including the LL and GRE models, as well as the finite mixture extended models including FSFM, LLFM and GREFM models, all produce comparable parameter estimates, which are also similar to the true values. More complex models yield better goodness of fit to the data with smaller deviances and smaller differences between the observed and expected frequencies (see Table 3 in Appendix A). In addition, all six models have very high matching accuracy with an AUC well above 99% and their ROC curves are indistinguishable from each other.

4.3. *Summary of results*. Incorporating conditional dependence in the two blocking schemes of the hospital linkage example shows comparable or improved matching accuracy relative to that of the FS model. In the SSN blocking scheme, conditional dependence models produce much improved model fit and higher matching accuracy. In the LNFN blocking scheme, conditional dependence models provide improved model fit and comparable matching accuracy. Comparison of data from the two blocking schemes reveals that fields used for matching in the SSN blocking scheme have relatively poor discriminating power and there exists strong conditional dependence in the dominating class. The ratios of m- and u-probabilities, known as the positive likelihood ratios, are lower than 10 for all fields in the SSN blocking scheme. In comparison, the m- and u-probability ratios are well above 10 into hundreds and even thousands for some fields in the LNFN blocking scheme, indicating extremely discriminating fields. The high field discrimination leads to high matching accuracy for all latent class models, even for the FS model whose conditional independence assumption is inappropriate.

As mentioned in Section 2, 7.6% and 4% of the manual review sample in the SSN and LNFN blocking schemes respectively had inconsistent reviews. A third adjudicator was employed to determine the match status. In order to examine whether these inconsistent reviews impact the results, three additional analyses were performed, where match status of record pairs with inconsistent reviews were determined differently. In Scenario 1, we assumed that the match status determined by the adjudicator was incorrect for all such record pairs. In Scenario 2, all these record pairs were assumed to be nonmatches. In Scenario 3, all these record pairs were assumed to be matches. ROC curves for each latent class model against true match status determined under the three scenarios are shown in Figure 4 in Appendix A. Based on these figures, we could again see that conditional dependence models produce improved matching accuracy compared to the FS model for the SSN blocking scheme, while all models yield comparable matching accuracy for the LNFN blocking scheme. This indicates that inconsistent reviews had little impact on the results.

In addition to the hospital linkage example, two additional real-world data linkage examples presented in Appendix B and Appendix C yield similar findings. In these two examples, accommodating conditional dependence also results in comparable or improved matching accuracy relative to that of the FS model. Utilization of poorly discriminating fields results in suboptimal matching performance of the FS model, which is improved by accommodating conditional dependence. When highly discriminating fields are used for matching, all models yield similar matching accuracy.

Based on these results, we hypothesize that incorporating the conditional dependence enables us to recover the true model parameters and enhance the matching accuracy in situations where matching fields have poor discriminating power. When the discriminating power of matching fields is high, incorporating the conditional dependence may not further improve matching accuracy relative to the FS model, despite the more satisfactory model fit. These hypotheses will be evaluated using a simulation study, where multiple factors including the discriminating power of matching fields are examined for their impact on whether and when it is important to accommodate conditional dependence.

5. Simulation study. The real-world examples in Section 4, Appendix B and Appendix C show that conditional dependence latent class models provide better fit to the data and yield matching accuracies at least as good as or better than that of the FS model assuming conditional independence. In this section, we perform a simulation study to examine when conditional dependence models will provide improved matching accuracy and what factors might play a role. Data will be simulated to represent the agreement or disagreement of the matching fields. These data can be thought of as vector patterns obtained after applying a specific blocking scheme in the real applications and are used directly in the latent class analysis. We will consider four scenarios in our simulation. Scenarios I, II and III have six fields with moderate power for discriminating matches from nonmatches (m-probabilities are 0.45, 0.25, 0.85, 0.95, 0.98, 0.99, and u-probabilities are 0.2, 0.05, 0.2, 0.3, 0.1, 0.05), but scenario I has field dependence in both match and nonmatch classes, scenario II has dependence only in the match class, whereas scenario III has dependence only in the nonmatch class. Scenario IV has fields with greater discriminating power compared to the first three scenarios (m-probabilities are 0.85, 0.9, 0.85, 0.95, 0.98, 0.99, and u-probabilities are 0.05, 0.1, 0.02, 0.05, 0.02, 0.01), with conditional dependence in both match and nonmatch classes.

For each scenario, we consider two values for the sample size ($N = 500{,}000$ and 5000 record pairs) and nine values for the true match prevalence (2%, 5%, 10%, 30%, 50%, 70%, 90%, 95%, 98%). For each combination of scenario, sample size, and true match prevalence, we generate 500 Monte Carlo data sets based on the finite mixture extended GRE model with $\eta_1 = 0.05$ and $\eta_0 = 0.2$, which are chosen based on the results of the hospital linkage example in Section 4. For all scenarios, we generate conditional dependence among the first four fields. For scenario I, the first four fields are correlated with $b_{j1} = 1, 2, 0.5, 1.5$ among true matches and $b_{j0} = 0.5, 1, 2, 0.5$ among true nonmatches. This results in a relatively stronger field dependence in the match class than in the nonmatch class, with tetrachoric correlation ranging from 0.32 to 0.74 among matches and from 0.2 to 0.63 among nonmatches. For scenario II, we generate the same dependence in the match class as in Scenario I but no dependence in the nonmatch class. For scenario III, we generate the same dependence in the true nonmatch class as in Scenario I but no dependence in the match class. Scenario IV is set up to have the same dependence in both match and non-match classes as in Scenario I.

For each simulated data set, we fit the FS, FSFM, LL, LLFM, GRE and GREFM models. Due to the structured field dependence of the simulated data, we include the pairwise interactions only among the first four fields in the LL and LLFM models, and assume nonzero b_{jd}'s only for the first four fields in the GRE and GREFM models. The correlations are considered in both match and nonmatch classes for both scenarios I and IV, in the match class only for scenario II and in the nonmatch class only for scenario III. Among the six models examined, the LLFM and GREFM models are the only two that can adequately account for the field dependence. We summarize estimates of match prevalence, m- and u-probabilities, and matching accuracy in terms of AUC across the 500 replicates. The average AUCs are shown in Figure 3. The biases in estimated prevalence and m- and u-probabilities are presented in the figure in Supplementary Material Xu et al. (2019). The magnitude of biases is represented using blue (under-estimation) to red (over-estimation) spectrum with lighter color for smaller bias and darker color for larger bias.

For scenario I with $N = 500{,}000$, none of the models produce bias when the true match prevalence is close to 50%. When the true match prevalence deviates from 50% in either direction, models that do not address the conditional dependence adequately start to show bias and the magnitude of the bias becomes greater as the true match prevalence moves further away from 50%. This is particularly true for the FS model. When the true match prevalence is close to 0%, the match prevalence is over-estimated while the m- and u-probabilities are generally under-estimated. On the other hand, the match prevalence is under-estimated and m- and u-probabilities are generally over-estimated when the true match prevalence approaches 100%. In terms of the matching accuracy, when the true match prevalence is near 50%, all six models produce indistinguishable ROC curves with similar AUCs of approximately 99.9% (Figure 3 panel A). These models continue to have similar matching accuracy when the true match prevalence deviates from 50% until it approaches either 0% or 100%. A difference in matching accuracy is observed with the FS and FSFM models and is most notable when the true match prevalence is less than 10% or greater than 90%. Specifically, the FS model is less accurate than the LL, LLFM, GRE and GREFM models when the true match prevalence is 2%, 5%, 10% and 98%, and the FSFM model is less accurate when the true match prevalence is 2%, 5% or 98%, with a greater difference at more extreme match

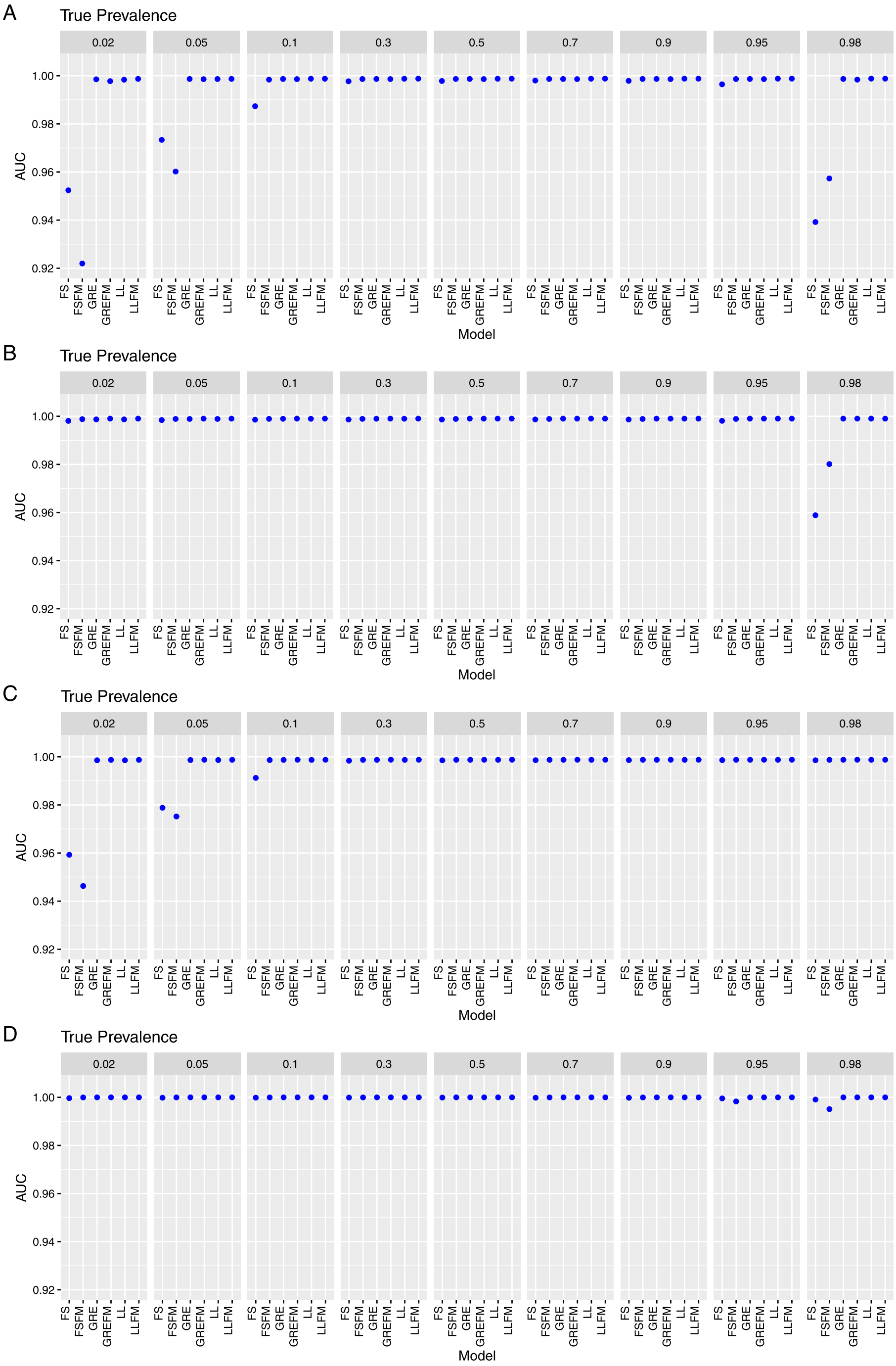

FIG. 3. *The average AUC (bottom panel) over* 500 *simulated data sets for latent class models in Scenario I (panel A), II (panel B), III (panel C) and IV (panel D) with* $N = 500{,}000$ *record pairs.*

prevalences. It is worth noting that the FSFM model generally gives the same or better accuracy compared to the FS model; however, at 2% prevalence it produces lower accuracy than the FS model. These results show that the FS model provides reasonable matching accuracy when the true match prevalence is close to 50%. Consideration of the conditional dependence models is important in the more extreme situations when the majority of the data are matches or nonmatches but the importance diminishes as the match prevalence is around 50%.

In order to examine how the conditional dependence in each class affects the model performance, we compare scenario I to scenarios II and III. In scenario II where fields are correlated only in the match class, all models yield similar matching accuracy close to 100%, except at the very high match prevalence of 98% when data are dominated by true matches (Figure 3 panels B and C). Specifically, at 98% prevalence, ignoring the conditional dependence produces biased parameter estimates and compromised matching accuracy. The FSFM model provides relatively better matching accuracy than the FS model, but the accuracy is not as good as the LL and GRE models and their finite mixture extensions. In scenario III where fields are correlated only in the nonmatch class, ignoring the conditional dependence produces biased parameter estimates when data are predominantly nonmatches with a true match prevalence 2–10%. The LL, LLFM, GRE and GREFM models all provide accurate matching regardless of the true match prevalence. The FS and FSFM models provide similar matching accuracy except when the true match prevalence decreases to 2–10%. These findings demonstrate the importance of addressing the conditional dependence when it exists in the vast majority of the data. In other words, we need to consider conditional dependence models if the conditional dependence exists in the dominating class of the data. On the other hand, if the conditional dependence exists only in the nondominating class, fitting conditional dependence models does not substantially improve the matching accuracy over the FS model.

We now look at the scenario IV to examine whether findings are similar when fields with higher discriminating power are used for record linkage. We again see that, when the true match prevalence is close to 50%, none of the models produces bias. When the true match prevalence approaches 0% or 100%, bias starts to appear for the FS and FSFM models. However, the magnitude of the bias is less severe compared to scenario I, especially for the prevalence estimate. Among all parameters, the u-probabilities of the FS and FSFM models show the greatest amount of bias when the true match prevalence is 95%–98%. However, these biases do not result in compromised matching accuracy as all models show similar AUCs (Figure 3 panel D). With a match prevalence close to 100%, nonmatches only constitute a small portion of the data. Hence bias in the u-probabilities only affects a small percentage of the data and does not have much impact on the overall matching accuracy.

Note that the simulation results shown above only present the averages across 500 simulated data sets because the variability of the parameter estimates and

AUCs are quite small with standard deviations in the order of 10^{-3}. In addition, simulation results for the smaller sample size of $N = 5000$ record pairs are not presented because they show the same results except that the parameter and AUC estimates exhibit greater variability. Also note that we only consider situations with moderately and highly discriminating fields in our simulations because record linkage practitioners would choose such fields from a set of candidates to link records. In situations where fields have poor discriminating power, estimates of match prevalence and m- and u-parameters based on the FS model are biased even when the true match prevalence is close to 50%, although the bias is not as severe as in situations when the true match prevalence is close to 0% or 100% (based on additional simulations, results not shown in manuscript but available from http://pages.iu.edu/~huipxu/publications.html). The FS model always provides inferior matching accuracy relative to the conditional dependence models, regardless of the true match prevalence. In our specific simulation, the AUCs of the FS model range from 5% lower when the true match prevalence is close to 50% to 15% lower when the true prevalence is close to 0% or 100%.

6. Discussion. The FS model is widely used in probabilistic record matching, despite its often invalid assumption of conditional independence. Prior literature has recognized the limitations of the FS model—biased parameter estimation when the conditional independence assumption fails. However, little investigation has been performed to evaluate the extent to which the record matching accuracy is impacted by the assumption. In this paper, we apply latent class models to the motivating hospital linkage example, with conditional dependence structure informed by the true match status of manually reviewed record pairs. In the SSN blocking scheme where fields have poor discriminating power, conditional dependence models yield improved matching accuracy compared to the FS model. In the LNFN blocking scheme where fields with good discriminating power are used for matching, incorporating conditional dependence results in comparable matching accuracy relative to the FS model. These findings are confirmed by the extensive simulation study, demonstrating that models incorporating the correct conditional dependence yield matching accuracies as good as or better than that of the FS model. In some situations, the simple FS model performs similarly to models with more complex conditional dependence structures. However, it is important to note that addressing conditional dependence is important when the true match prevalence approaches 0% or 100% and the conditional dependence exists in the dominating class. When conditional dependence lies in the nondominating class only or when the match prevalence is near 50%, it becomes less important to consider conditional dependence models as all models produce comparable matching accuracy. Note that one may still prefer conditional dependence models in cases with extremely large sample sizes, where a slight improvement in matching accuracy could translate to a large number of record pairs being correctly classified.

Our study further shows that the need to address conditional dependence at a match prevalence close to 0% or 100% diminishes if the discriminating power of matching fields is high, as seen in the LNFN blocking scheme of the hospital linkage example. This finding is consistent with Fellegi and Sunter's claim in their seminal work, where they stated that they believe the FS model is robust to departures from the conditional independence assumption if sufficient identifying information is used for linkage operation (Fellegi and Sunter (1969)). When fields have high discriminating power, the FS model can provide a matching accuracy comparable to the better fitting conditional dependence models. In the literature, it has been well recognized that the gain in predictive accuracy from building increasingly more complex models decreases dramatically and simple models can account for over 90% of the achievable predictive power in many situations (Hand (2006)). In supervised classifications, the simple naive Bayes rule with conditional independence assumption can often perform surprisingly well and may even have better performance than complex rules (Hand and Yu (2001)). These authors commented that one of the reasons is the low variance in the probability estimates of the naive Bayes rule. Although the naive Bayes rule produces biased probability estimates, the bias may be inconsequential for classification as long as the rank order is preserved. The same arguments can be made with regard to the often satisfactory performance of the FS model for the unsupervised classification of record linkage problems.

In contrast, if fields have poor discriminating power, as seen in the SSN blocking scheme of the hospital linkage example, it is important to consider conditional dependence even if the true match prevalence is close to 50%. This is due to the inferior matching accuracy of the FS model, regardless of the true match prevalence. These findings highlight the importance of using highly discriminating fields for record linkage whenever possible. These results are also consistent with previous research that assessed the discriminating power of matching fields using real-world record linkage problems (Cook, Olson and Dean (2001), Quantin et al. (2004)). In particular, our work adds to the literature that fields with high discriminating power are critical especially in situations when the match prevalence is extreme and dependence exists in the dominating class. In addition, these findings explain why the FS model produces matching accuracy inferior to that of conditional dependence models only in the SSN blocking scheme of the hospital linkage example but comparable matching accuracy in the LNFN blocking scheme: The SSN blocking scheme uses fields with poor discriminating power where strong correlation exists in the class that makes up 97% of the record pairs. The LNFN blocking scheme, on the other hand, uses highly discriminating fields. Hence all models produce comparable results. This also explains why, in Daggy et al. (Daggy et al. (2014)), the FS model produced a negligible misclassification rate at low match prevalence even when strong conditional dependence existed in both match and nonmatch classes in their simulation study: Several fields in the simulations were extremely discriminating with u-probabilities of 0.001 and m-probabilities of 0.95. Had these u-probabilities been larger and m-probabilities been smaller, the misclassification rate of the FS model would have been much higher.

Lastly, we caution that not all conditional dependence models will necessarily improve the matching accuracy when conditional dependence exists. The finite mixture extended FS model addresses the conditional dependence; however, in our simulations, it sometimes produces less accurate record matching than the FS model, despite a better fit to the data. This finding suggests that a conditional dependence model with misspecified dependence structure could potentially produce inferior matching accuracy than the FS model, consistent with earlier findings on the impact of incorrect conditional dependence structures for record linkage (Li et al. (2018)). This is also consistent with findings by Albert and Dodd (2004) regarding diagnostic test accuracy evaluation—conditional dependence models lack robustness and thus parameter estimates can be seriously biased with misspecified conditional dependence structure. We therefore caution against a blind use of conditional dependence models, even if there is a need to address the conditional dependence. Studies have successfully used multiple diagnostic tools to identify the conditional dependence (Garrett and Zeger (2000), Qu, Tan and Kutner (1996), Sepúlveda, Vicente-Villardón and Galindo (2008)). However, recent research has found that these approaches may not be able to identify the appropriate dependence structure in situations when the conditional independence model is highly biased (Subtil, de Oliveira and Gonçalves (2012)). These authors further demonstrated that the dependence would be correctly identified if the true values of parameters were used to estimate the expected correlations. Moreover, to decide whether to address conditional dependence in record linkage, one needs to have an approximate estimate of the true match prevalence. This can be difficult to obtain since the FS model can give severely biased estimate of the match prevalence in some situations. Owing to these issues, we recommend human review of a sample of record pairs to establish the true match status whenever feasible for practical applications. This will help in both choosing a model with appropriate dependence structure and establishing an estimate for match prevalence. If human review is not available, a sensitivity analysis should be conducted to apply latent class models with various structures to address the conditional dependence.

The finite mixture extension of the conditional dependence models appears to perform similarly to the corresponding models without the finite mixture extension in simulation studies in terms of both parameter estimates and matching accuracy. This is due to the relatively low percentages of record pairs whose field agreement pattern is always consistent with the underlying true match status, which were set to be 5% in the match class and 20% in the nonmatch class. When these percentages are higher, additional simulations (results not shown but reproducible with a SAS program available at http://pages.iu.edu/~huipxu/publications.html) show that the conditional dependence models with the finite mixture extension provide less biased parameter estimates than models without the finite mixture extension, but the matching accuracy is still comparable. This again confirms our previous finding that improved model fit does not necessarily lead to improved matching accuracy. Furthermore, comparison between LL models and GRE models shows that both models produce comparable results. However, as reported by Daggy et al. (Daggy et al. (2014)), GRE models are computationally more intensive. They are also more likely to have numerical instability. In practice, GRE models can be a useful tool when the conditional dependence is prevalent among many fields as they involve fewer parameters than LL models. We recommend using multiple starting values if GRE models are used since the convergence of these models is strongly dependent on the starting values.

APPENDIX A

This section shows the model estimated frequencies for each vector pattern of the SSN and LNFN blocking schemes of the hospital linkage data. Also shown are the ROC curves of the latent class models where the true match status of record pairs with inconsistent manual reviews was assumed under three scenarios.

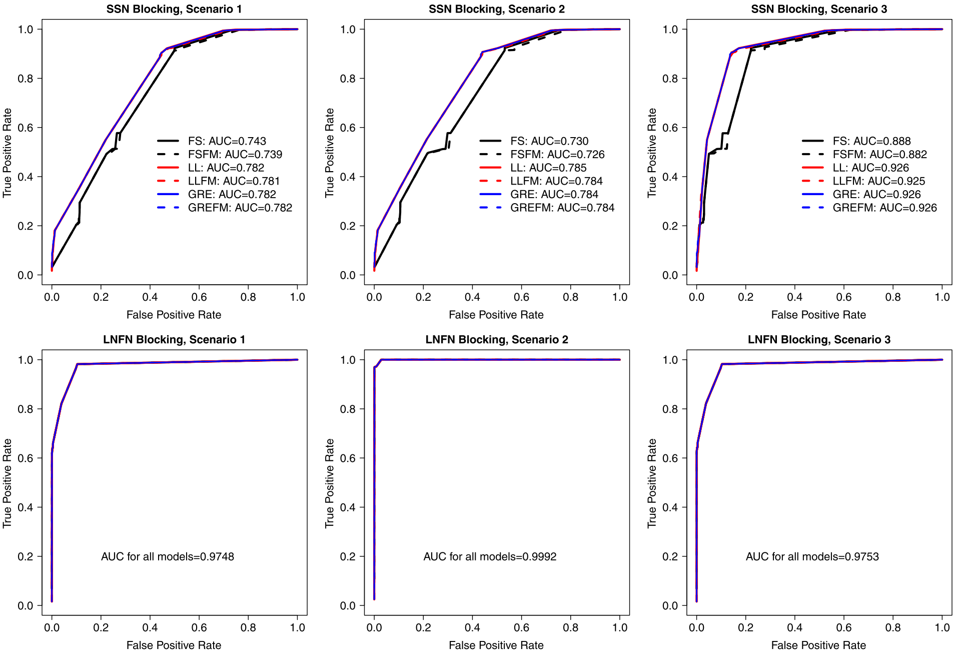

FIG. 4. *ROC curves of latent class models of the hospital linkage data, where the match status of record pairs with discrepant evaluation between two reviewers was derived assuming that the third adjudicator was incorrect* (*Scenario* 1), *all were nonmatches* (*Scenario* 2) *and all were matches* (*Scenario* 3).

TABLE 2
Expected frequencies estimated by the latent class models for the SSN blocking scheme of the hospital linkage data

	Manual review		Total	Expected frequency					
Vector		Non-	fre-						
pattern	Match	match	quency	FS	FSFM	LL	LLFM	GRE	GREFM
00000	53	158	5345	1541.3	5345.0	5329.6	5345.0	5226.6	5345.0
00001	1811	190	48,981	50,619.6	47,035.5	48,984.5	48,975.9	49,238.7	48,979.6
00010	242	42	2898	5773.4	2830.2	2916.8	2910.7	2886.7	2877.3
00011	807	118	198,301	198,489.4	200,776.9	198,333.3	198,281.2	198,852.8	198,368.7
00100	4	2	92	240.3	111.8	103.1	102.4	122.0	112.2
00101	134	1	5831	7957.1	7764.7	5534.1	5809.2	5560.2	5783.8
00110	33	3	243	906.8	471.2	239.0	237.8	191.1	244.3
00111	61	6	36,498	32,324.6	33,792.1	36,748.5	36,526.9	36,074.6	36,480.4
01000	1	0	24	20.1	14.6	33.2	33.2	52.7	32.9
01001	17	3	1508	1370.2	1418.5	1735.3	1499.7	1357.3	1490.3
01010	21	1	144	148.3	155.6	98.7	132.1	125.7	123.9
01011	38	9	17,005	17,434.6	17,609.0	16,774.7	17,028.2	17,142.6	17,158.0
01100	0	0	3	3.7	2.9	1.3	1.1	1.7	1.8
01101	1	0	257	305.2	290.5	319.8	274.6	294.6	251.8
01110	3	0	16	32.7	33.3	18.1	16.8	20.6	16.9
01111	14	2	4352	4330.6	3846.4	4327.9	4323.3	4337.4	4232.0
10000	18	9	409	287.0	160.1	412.8	351.7	461.9	335.3
10001	256	44	13,218	12,875.9	13,059.7	13,428.1	13,305.6	12,872.0	13,360.3
10010	115	14	1195	1429.6	1123.4	1176.5	1194.2	1217.0	1262.9
10011	356	41	111,353	109,129.7	111,260.7	111,118.4	111,335.7	110,587.1	111,071.6
10100	1	0	13	47.4	28.6	11.0	17.3	12.6	12.7
10101	25	1	1872	2460.8	2426.6	1957.8	1854.0	2381.7	1854.3
10110	19	0	108	270.4	222.7	110.6	108.4	112.6	99.6
10111	37	0	23,019	25,022.7	22,967.0	22,971.8	23,020.1	23,559.7	23,188.9
11000	11	3	97	32.1	43.1	95.0	129.9	136.6	124.5
11001	86	10	5544	5039.0	5222.9	5107.6	5485.0	5113.2	5487.5
11010	76	12	736	529.7	698.2	754.6	740.2	720.5	729.7
11011	225	32	85,415	87,453.1	85,419.6	85,874.1	85,424.6	86,180.5	85,425.6
11100	0	0	4	8.1	9.6	4.4	3.6	6.2	4.8
11101	13	1	906	1296.3	1170.0	1046.4	915.9	1266.7	927.3
11110	13	1	80	136.2	156.8	102.4	82.7	112.6	83.3
11111	31	5	24,661	22,612.1	24,661.0	24,458.6	24,661.0	23,902.3	24,661.0

APPENDIX B

In this section, we will further evaluate whether incorporating conditional dependence might improve the matching accuracy using a publicly available data set for the deduplication of personal data records from the Epidemiological Can-

TABLE 3

Expected frequencies estimated by the latent class models for the LNFN blocking scheme of the hospital linkage data

Vector	Manual review		Total	Expected frequency					
pattern	Match	Nonmatch	frequency	FS	FSFM	LL	LLFM	GRE	GREFM
000000	0	482	11,905,306	11,881,099	11,905,306	11,889,364	11,905,306	11,889,456	11,905,306
000001	2	39	1,039,570	1,055,810	1,035,579	1,052,341	1,038,721	1,052,279	1,038,733
000010	0	6	134,229	140,017	134,703	135,919	131,238	135,889	131,227
000011	0	0	12,734	12,511	16,077	12,076	15,249	12,073	15,236
000100	0	0	1168	5136	5148	1271	1234	1252	1211
000101	0	2	118	483	632	120	148	117	145
000110	0	1	4302	61	80	4638	4499	4662	4514
000111	9	28	502	23	19	434	539	436	540
001000	0	16	382,209	394,884	380,910	392,160	382,915	392,113	382,921
001001	29	15	51,476	39,316	50,983	41,327	49,952	41,362	49,950
001010	0	1	4657	4684	5919	4502	5623	4501	5618
001011	6	1	3409	3210	3532	2233	2210	2233	2209
001100	0	1	54	182	234	45	55	44	54
001101	14	3	227	1080	750	275	238	250	216
001110	9	8	189	10	8	163	199	163	200
001111	51	5	881	704	357	906	788	912	793
010000	1	1	1704	3096	3020	2300	2437	2267	2410
010001	2	0	492	394	504	334	380	331	376
010010	6	0	799	37	48	1060	1150	1088	1168
010011	16	1	303	82	77	166	187	168	189
010100	2	0	48	2	2	2	3	38	53
010101	2	0	25	30	19	10	8	14	14
010110	14	3	190	0	0	317	383	304	373
010111	6	0	118	20	9	93	93	91	91
011000	1	0	169	155	196	130	144	129	142

TABLE 3
(*Continued*)

Vector	Manual review		Total	Expected frequency					
pattern	Match	Nonmatch	frequency	FS	FSFM	LL	LLFM	GRE	GREFM
011001	7	1	4113	4763	5253	5036	4320	5030	4313
011010	5	0	84	36	34	65	71	66	72
011011	3	0	2395	3141	2580	2776	2383	2776	2381
011100	0	0	8	13	8	4	3	6	5
011101	0	0	271	1196	659	381	328	404	347
011110	1	0	27	9	4	37	36	37	35
011111	0	0	1197	790	325	2490	2134	2487	2131
100000	3	12	216,212	224,110	215,230	221,982	217,269	221,952	217,284
100001	33	3	30,843	24,589	31,959	26,664	31,222	26,697	31,229
100010	0	0	3109	2675	3363	2558	3197	2557	3195
100011	20	2	2796	3327	3550	2171	2062	2169	2058
100100	0	0	86	110	137	27	32	26	31
100101	28	0	248	1186	821	290	255	263	231
100110	17	0	423	10	7	97	117	97	117
100111	62	0	918	778	397	954	844	959	847
101000	15	2	11,616	9514	12,222	10,234	11,746	10,245	11,744
101001	2589	10	271,855	188,050	231,342	271,294	268,289	272,525	269,489
101010	6	1	1069	1454	1536	892	826	890	823
101011	720	14	69,634	123,814	113,424	75,007	74,899	74,871	74,778
101100	2	0	90	524	361	120	104	109	94
101101	151	0	7525	47,149	28,966	11,090	11,118	10,027	10,072
101110	20	0	332	344	175	396	344	397	344
101111	357	0	40,208	31,151	14,268	36,447	36,758	36,625	36,930
110000	30	0	1117	116	146	100	102	100	101
110001	9	0	4072	5262	5843	5334	4682	5323	4666

TABLE 3
(*Continued*)

Vector pattern	Manual review		Total frequency	Expected frequency					
	Match	Nonmatch		FS	FSFM	LL	LLFM	GRE	GREFM
110010	14	0	620	39	36	51	52	52	52
110011	0	0	2394	3473	2874	2941	2583	2938	2577
110100	1	0	85	15	9	4	3	5	5
110101	1	0	298	1323	734	403	355	428	376
110110	9	0	490	10	4	34	31	34	30
110111	1	0	1376	874	362	2639	2315	2633	2307
111000	1	0	1385	2325	2574	2214	1909	2207	1899
111001	269	0	204,020	210,460	209,662	205,346	206,766	204,755	206,198
111010	0	0	850	1535	1267	1221	1053	1218	1048
111011	150	0	117,785	139,049	103,277	113,229	114,085	113,018	113,882
111100	0	0	93	584	323	167	145	177	153
111101	28	0	19,407	52,954	26,378	15,541	15,722	16,455	16,610
111110	0	0	450	386	159	1095	944	1092	939
111111	120	0	100,768	34,986	100,768	101,632	102,348	101,333	102,062

cer Registry of North Rhine-Westphalia in Germany (Schmidtmann et al. (2009)). The comparison patterns in the data set were formed based on a sample of 100,000 records collected between 2005 and 2008 and are available at https://archive.ics.uci.edu/ml/datasets/Record+Linkage+Comparison+Patterns. Due to the large number of possible pairs, six blocking schemes were utilized, resulting in 5,749,132 record pairs, of which 20,931 pairs were matches. The true match status was ascertained by applying two record linkage software packages, where record pairs classified as a match or a potential match by one or both software packages were subjected to an extensive manual review involving three experienced documentarists and four further staff members (Sariyar, Borg and Pommerening (2011)). Seven fields were available in the data set, including first name, last name, sex, birth day, birth month, birth year and postal code. The agreement of the last name and first name was measured as a value between 0 and 1, indicating the phonetic similarity of the names. Binary agreement in last name and first name was then derived based on the dichotomization with 0.9 as the threshold, while for the other five fields, exact agrement was derived.

In our analysis, we focus on two blocking schemes: one requiring agreement in date of birth (DOB) and the other requiring equality of last name and sex (LNSEX). The DOB blocking scheme involves 331,637 record pairs, of which 20,766 (6.26%) pairs are matches. The LNSEX block includes 732,897 record pairs, of which 20,463 (2.79%) pairs are matches. Examination of the true matches and true nonmatches reveals that last name and sex are correlated among the nonmatches in the DOB blocking scheme, while in the LNSEX blocking scheme, day, month and year of birth were correlated among the matches. We therefore fit the LL model with pairwise interactions between last name and sex in the nonmatch class for the DOB blocking scheme, allowing fields to be independent in the match class. The conditional dependence will also be modeled using the GRE model with nonzero $b_{jM} = b$ for last name and sex in the nonmatch class. For the LNSEX blocking scheme, conditional dependence will be accommodated using the LL model with pairwise interactions among birth day, month and year and using the GRE model with nonzero b_{jM} for these three fields in the match class only.

Model deviance and estimates of model parameters are shown in Table 4. For the DOB blocking scheme, accommodation of the conditional dependence, which lies in the nonmatching class or the dominating class, results in substantial improvement in model fit, as seen by remarkably smaller deviance of the conditional dependence models relative to the FS model. The estimated frequencies based on conditional dependence models are much closer to the observed frequencies, as shown in Table 5. Model parameter estimates, on the other hand, are very similar across models. In addition, all models provide excellent discrimination with an AUC above 99.9%.

For the LNSEX blocking scheme, accommodation of the conditional dependence, which lies in the nondominating class, results in better model fit than the

TABLE 4

Estimates of match prevalence and m- and u-probabilities of latent class models for the deduplication of Cancer Registry Data

	Manual review	FS	FSFM	LL	LLFM	GRE	GREFM
			The DOB blocking scheme				
Prevalence	0.0626	0.0625	0.0626	0.0625	0.0626	0.0625	0.0626
m-probabilities							
First name	0.9905	0.9915	0.9912	0.9915	0.9906	0.9915	0.9906
Last name	0.9922	0.9928	0.9927	0.9931	0.9923	0.9931	0.9923
Sex	0.9873	0.9876	0.9873	0.9874	0.9874	0.9874	0.9874
Postal code	0.9575	0.9583	0.9578	0.9585	0.9572	0.9585	0.9572
u-probabilities							
First name	0.0092	0.0092	0.0092	0.0092	0.0091	0.0092	0.0091
Last name	0.0004	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004
Sex	0.5048	0.5048	0.5048	0.5049	0.5048	0.5049	0.5048
Postal code	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016	0.0016
Deviance		3321.7	882.7	36.1	0.9	36.1	0.9
			The LNSEX blocking scheme				
Prevalence	0.0279	0.0279	0.0280	0.0279	0.0280	0.0279	0.0280
m-probabilities							
First name	0.9910	0.9915	0.9898	0.9915	0.9899	0.9915	0.9898
Birth day	0.9968	0.9979	0.9973	0.9959	0.9951	0.9971	0.9965
Birth month	0.9977	0.9986	0.9982	0.9968	0.9962	0.9979	0.9976
Birth year	0.9961	0.9974	0.9967	0.9956	0.9948	0.9968	0.9962
Postal code	0.9585	0.9592	0.9568	0.9592	0.9566	0.9592	0.9567
u-probabilities							
First name	0.0145	0.0145	0.0145	0.0145	0.0144	0.0145	0.0145
Birth day	0.0327	0.0327	0.0327	0.0327	0.0327	0.0327	0.0327
Birth month	0.0824	0.0824	0.0824	0.0824	0.0824	0.0824	0.0824
Birth year	0.0239	0.0239	0.0238	0.0239	0.0238	0.0239	0.0238
Postal code	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035
Deviance		119.18	84.38	79.78	54.38	98.18	67.98

FS model. However, the improvement is rather modest. Estimated parameter values of all six models are similar to the true values and all models yield excellent discrimination with an AUC above 99.9%.

These results show that accommodating conditional dependence leads to comparable matching accuracy relative to that of the FS model while improving model fit. Both blocking schemes use fields such as first name, last name and postal codes with high discriminating power, resulting in high matching accuracy for all latent class models, even for the FS model whose conditional independence assumption is inappropriate.

TABLE 5

Expected frequencies estimated by the latent class models for cancer registry data deduplication

Vector	Observed frequency		Expected frequency					
pattern	Matches	Total	FS	FSFM	LL	LLFM	GRE	GREFM
			The DOB blocking scheme					
0000	0	153,560	152,211.9	153,560	153,551.3	153,560	153,551.3	153,560
0001	1	252	249.74	54.39	251.77	250.04	251.77	250.04
0010	0	153,843	155,189.4	153,843	153,851.5	153,843	153,851.5	153,843
0011	11	253	255.8	451.53	253.4	254.95	253.4	254.95
0100	2	60	68.99	14.85	69	61.95	69	61.95
0101	3	4	2.18	5.11	2.22	5.98	2.22	5.98
0110	22	86	77.44	136.11	76.2	83.34	76.2	83.34
0111	159	163	165.5	156.03	165.6	161.69	165.6	161.69
1000	0	66	1409.19	307.01	66.25	66	66.25	66
1001	5	5	4.06	5.08	1.82	4.9	1.82	4.9
1010	13	2790	1442.76	2547.96	2789.87	2789.99	2789.87	2789.99
1011	133	136	142.1	136.05	139.09	136.08	139.09	136.08
1100	31	31	11.15	26.76	10.67	29.77	10.67	29.77
1101	221	221	241.76	225.8	245.94	220.36	245.94	220.36
1110	814	816	841.12	816.33	837.21	817.9	837.21	817.9
1111	19,351	19,351	19,323.87	19,351	19,325.15	19,351	19,325.15	19,351
			The LNSEX blocking scheme					
00000	0	606,312	606,116.49	606,311.99	606,155.4	606,311.9	606,139.9	606,312
00001	0	2132	2157.01	2145.87	2122.59	2117.13	2141.65	2136.03
00010	0	14,747	14,825.63	14,751.2	14,822.81	14,761.26	14,821.8	14,753.9
00011	0	46	52.76	53.33	51.95	52.7	52.44	53.28
00100	0	54,422	54,455.17	54,341.53	54,457.6	54,368.3	54,453.87	54,353.4
00101	0	180	193.79	196.45	190.72	192.95	192.46	195.18
00110	0	1277	1331.99	1350.73	1331.71	1343.88	1331.56	1346.48
00111	0	4	5.09	6.93	5	6.46	5.04	6.67
01000	0	20,483	20,514.38	20,428.05	20,515.85	20,446.36	20,514.47	20,437.3
01001	0	78	73.01	73.86	71.85	72.61	72.51	73.39
01010	0	502	501.79	507.91	501.7	505.5	501.64	506.41
01011	2	2	2.02	3.17	1.98	2.86	1.99	3.01
01100	1	1856	1843.09	1870.46	1843.18	1861.42	1842.98	1865.12
01101	1	9	6.99	9.26	6.88	8.77	6.93	9.05
01110	22	86	52.15	87.39	52.15	83.8	52.14	86.01
01111	158	162	166.64	159.88	166.66	162	166.65	160.74
10000	0	8775	8933.71	8889.07	8899.11	8864.69	8917.99	8880.12
10001	14	62	31.79	32.13	60.14	55.88	35.38	34.23
10010	0	312	218.52	220.95	217.85	220.97	218.4	221.84
10011	4	6	0.83	1.14	7.03	9.6	8.61	8.64
10100	0	837	802.63	813.79	799.63	808.13	801.45	811.68
10101	4	7	2.96	3.58	8.26	10.71	9.87	9.3
10110	5	32	21.34	30.62	21.2	28.51	21.22	29.44

TABLE 5
(*Continued*)

Vector pattern	Observed frequency		Expected frequency					
	Matches	Total	FS	FSFM	LL	LLFM	GRE	GREFM
10111	39	39	40.27	40.68	39.02	36.83	38.52	37.87
11000	1	269	302.37	305.97	301.25	304.17	301.96	305.21
11001	1	2	1.15	1.52	3.22	5.13	4.48	3.56
11010	0	13	8.54	14.37	8.46	12.91	8.45	13.6
11011	26	26	27	26.46	25.83	23.22	25.31	24.46
11100	9	41	29.29	40.72	29.14	38.59	29.19	39.82
11101	49	49	50.17	49.75	49.21	47.15	48.94	48.58
11110	812	814	820.9	813.24	820.98	817.49	820.98	815.64
11111	19,315	19,315	19,307.52	19,315	19,308.67	19,315.06	19,308.22	19,315

APPENDIX C

Another publicly available data set will be used to evaluate whether incorporating conditional dependence might improve the matching accuracy. This data set involves the disambiguation of inventor records in the United States Patent and Trademark Office (USPTO) database. A total of 98,762 inventor records corresponding to a sample of 824 inventors in the optoelectronics industry were manually disambiguated, where a random sample of 150,000 record pairs was selected from all pairwise comparisons and used as the training data for the classification models by Ventura, Nugent and Fuchs (2015) (available at http://www.cmu.edu/epp/disambiguation). There were eight fields in the data, where agreement for last name, first name, and middle initial of the inventor, city, and assignee was measured using the similarity scores based on the Jaro–Winkler string comparator. Binary agreement status was then derived according to the dichotomization of the similarity scores at the 0.8 threshold.

Among the 150,000 record pairs, 19,896 pairs are true matches, resulting in a match prevalence of 13.3%. Further examination of the field agreement status among true matches and true nonmatches determined by manual review reveals that first name and last name show a strong negative correlation in the nonmatch class. Specifically, approximately 92% of the true nonmatches disagree on last name and agree on first name. Another 7.5% of the record pairs agree on last name while disagree on first name. No record pairs have disagreement on both last and first names. We therefore perform two analyses, both of which include first name and last name as matching fields. Three additional fields are included. In Analysis I, we include country, state and assignee that have relatively poor discriminating power. In Analysis II, we include middle name, city and suffix with good discriminating power. The strong conditional dependence between first name and last name

TABLE 6

Estimates of match prevalence and m- and u-probabilities of latent class models for the USPTO inventor records disambiguation data

	Manual review	FS	FSFM	LL	LLFM
		Analysis I			
Prevalence	0.1326	0.2012	0.2012	0.1367	0.1367
m-probabilities					
First name	1.0000	0.6748	0.6748	0.9922	0.9922
Last name	1.0000	1.0000	1.0000	0.9304	0.9304
Country	0.9932	0.9723	0.9723	1.0000	1.0000
Assignee	0.8371	0.6003	0.6003	0.8678	0.8678
State	0.9294	0.6377	0.6377	0.9756	0.9756
u-probabilities					
First name	0.9246	1.0000	1.0000	0.9254	0.9254
Last name	0.0791	0.0000	0.0000	0.0858	0.0858
Country	0.9548	0.9568	0.9568	0.9536	0.9536
Assignee	0.1204	0.1185	0.1185	0.1122	0.1122
State	0.0665	0.0659	0.0659	0.0551	0.0551
Deviance		41,224.4	7687.2	1528.2	1528.2
		Analysis II			
Prevalence	0.1326	0.1290	0.1290	0.1358	0.1358
m-probabilities					
First name	1.0000	0.9878	0.9878	1.0000	1.0000
Last name	1.0000	0.9989	0.9989	1.0000	1.0000
Middle name	0.8495	0.8597	0.8597	0.8308	0.8308
City	0.8055	0.8288	0.8288	0.7873	0.7873
Suffix	0.1945	0.2002	0.2002	0.1899	0.1899
u-probabilities					
First name	0.9246	0.9267	0.9267	0.9243	0.9243
Last name	0.0791	0.0831	0.0831	0.0757	0.0757
Middle name	0.0450	0.0468	0.0468	0.0450	0.0450
City	0.0083	0.0082	0.0082	0.0083	0.0083
Suffix	0.0019	0.0018	0.0018	0.0019	0.0019
Deviance		65,146.1	65,146.1	1716.5	1716.5

in the nonmatch class is accommodated using the LL model for both analyses. The GRE model is not used due to the numerical instability since the tetrachoric correlation underlying the binary agreement status is nearly on the boundary.

The match prevalence and the m- and u-probabilities estimated by latent class models are shown in Table 6. The LL model provides substantially better fit than the FS model for both analyses, with expected frequencies generally closer to the observed frequencies (Table 7). In Analysis I where the three additional fields have poor discrimination, parameter estimates are much less biased in the LL model

TABLE 7
Expected frequencies estimated by the latent class models for the USPTO inventor records disambiguation data

Vector pattern	Observed frequency		Expected frequency			
	Matches	Total	FS	FSFM	LL	LLFM
			Analysis I			
00000	0	0	327.48	0	0	0
00001	0	0	19.12	0	0	0
00010	0	0	42.52	0	0	0
00011	0	0	2.48	0	0	0
00100	0	0	6752.06	0	0.04	0.04
00101	0	0	399.03	0	1.44	1.44
00110	0	0	876.72	0	0.24	0.24
00111	0	0	83.48	0	9.47	9.47
01000	0	574	32.23	368.7	376.18	376.18
01001	0	0	1.88	114.09	21.95	21.95
01010	0	91	4.18	68.94	47.55	47.55
01011	0	0	0.24	21.33	2.78	2.78
01100	0	7313	664.46	5946.77	7723.31	7723.31
01101	0	525	101.05	1840.08	470.02	470.02
01110	0	1117	86.28	1112	979.35	979.35
01111	0	196	422.79	344.08	183.66	183.66
10000	0	4793	4251.93	4264.46	4612.68	4612.68
10001	0	0	248.25	300.7	269.2	269.2
10010	0	386	552.09	573.41	583.05	583.05
10011	0	0	32.23	40.43	34.03	34.03
10100	0	94,829	87,666.04	94,392.93	94,699.85	94,699.84
10101	0	5992	5299.33	6655.91	5709.06	5709.06
10110	0	11,915	11,383.03	12,692.2	11,999.67	11,999.67
10111	0	1900	1879.2	894.96	1896.28	1896.28
11000	7	33	418.42	169.17	56.69	56.69
11001	0	0	24.43	52.35	3.31	3.31
11010	129	138	54.33	31.63	7.17	7.17
11011	0	0	3.17	9.79	0.42	0.42
11100	754	1114	8627.03	2728.56	1224.89	1224.89
11101	2480	2505	2845.27	844.28	2509.39	2509.39
11110	515	556	1120.18	510.22	548.36	548.36
11111	16,011	16,023	15,779.06	16,023	16,029.98	16,029.98
			Analysis II			
00000	0	0	8289.56	8289.56	0	0
00001	0	0	15.3	15.3	0	0
00010	0	0	68.39	68.39	0	0
00011	0	0	0.13	0.13	0	0
00100	0	0	407.3	407.3	0	0
00101	0	0	0.76	0.76	0	0
00110	0	0	3.5	3.5	0	0

TABLE 7
(*Continued*)

Vector	Observed frequency		Expected frequency			
pattern	Matches	Total	FS	FSFM	LL	LLFM
00111	0	0	0.04	0.04	0	0
01000	0	9501	755.67	755.67	9279.29	9279.29
01001	0	13	2.52	2.52	17.64	17.64
01010	0	112	28.04	28.04	77.23	77.23
01011	0	0	5.48	5.48	0.15	0.15
01100	0	183	64.57	64.57	437.21	437.21
01101	0	0	6.99	6.99	0.83	0.83
01110	0	7	134.21	134.21	3.64	3.64
01111	0	0	33.52	33.52	0.01	0.01
10000	0	113,064	104,752.38	104,752.38	113,263.87	113,263.86
10001	0	223	193.37	193.37	215.35	215.35
10010	0	880	865.88	865.88	942.68	942.68
10011	0	5	2.08	2.08	1.79	1.79
10100	0	5572	5148.93	5148.93	5336.66	5336.66
10101	0	5	10.11	10.11	10.15	10.15
10110	0	66	54.24	54.24	44.42	44.42
10111	0	0	3.03	3.03	0.08	0.08
11000	364	810	9858.85	9858.85	593.98	593.98
11001	0	0	109.33	109.33	139.28	139.28
11010	2516	2522	1853.46	1853.46	2198.27	2198.28
11011	115	115	444.54	444.54	515.46	515.46
11100	2207	2224	2714.65	2714.65	2915.98	2915.98
11101	1299	1299	563.7	563.7	683.75	683.75
11110	10,940	10,944	10,885.35	10,885.35	10,791.76	10,791.76
11111	2455	2455	2724.09	2724.09	2530.5	2530.5

than in the FS model. In Analysis II where the additional fields have good discrimination, all models produce similar parameter estimates with little bias. In terms of matching accuracy, the AUC was 96% for the FS model and 99% for the LL model for Analysis I, implying lower accuracy of the FS model (see Figure 5). In Analysis II, however, the AUC was 99.9% for all models, suggesting little difference in matching accuracy.

Accommodating conditional dependence in this example results in comparable or improved matching accuracy relative to that of the FS model while improving model fit. Specifically, conditional dependence models produce higher matching accuracy in Analysis I, where fields used for matching have relatively poor discriminating power and there exists strong conditional dependence in the dominating class. The ratios of the m- and u-probabilities for the three additional fields are 1, 7 and 14 for country, assignee and state. In comparison, all latent class models produce comparable matching accuracy in Analysis II, where fields have much

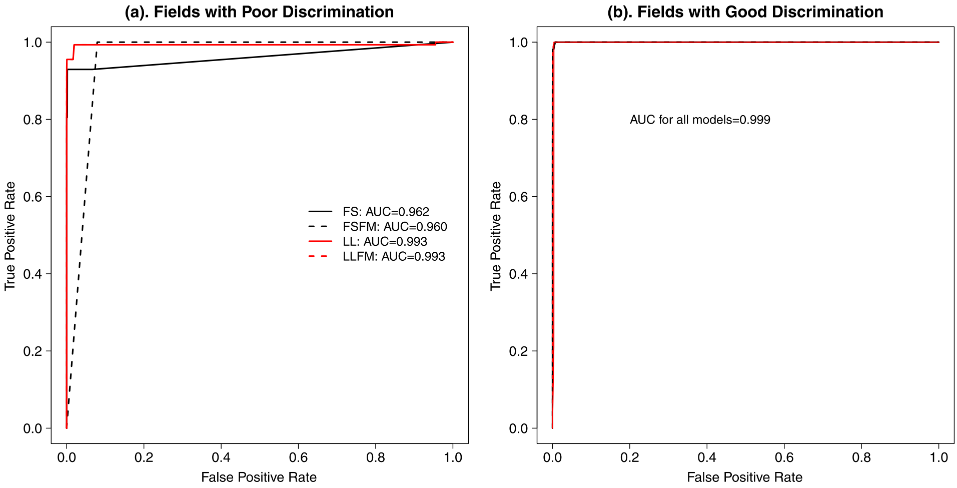

FIG. 5. *ROC curves of latent class models of the USPTO inventor records disambiguation data.*

greater discrimination, with ratios of the m- and u-probabilities being 19, 97 and 102 for middle name, city and suffix. These highly discriminating fields lead to high matching accuracy for all latent class models.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/19-AOAS1256SUPP; .pdf). The supplement includes an additional figure, presenting the average biases of the estimated match prevalence, m- and u-probabilities in the simulation study using a heat map.

REFERENCES

ABRIL, D., NAVARRO-ARRIBAS, G. and TORRA, V. (2012). Improving record linkage with supervised learning for disclosure risk assessment. *Inf. Fusion* **13** 274–284.

ALBERT, P. S. (2009). Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. *Stat. Med.* **28** 780–797. MR2657042

ALBERT, P. S. and DODD, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60** 427–435. MR2066277

ALBERT, P. S. and DODD, L. E. (2008). On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *J. Amer. Statist. Assoc.* **103** 61–73. MR2420217

ALBERT, P. S., MCSHANE, L. M. and SHIH, J. H. (2001). Latent class modelling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* **57** 610–619. MR1855699

ARMSTRONG, J. B. and MAYDA, J. E. (1992). Estimation of record linkage models using dependent data. In *Proceedings of the Section on Survey Research Methods* 853–858. Amer. Statist. Assoc., Providence, RI.

BILENKO, M. and MOONEY, R. J. (2003a). Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the* 9*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 39–48. ACM, New York.

BILENKO, M. and MOONEY, R. J. (2003b). On evaluation and training-set construction for duplicate detection. In *Proceedings of the KDD*-03 *Workshop on Data Cleaning*, *Record Linkage*, *and Object Consolidation* 7–12. Washington, DC.

CHRISTEN, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 151–159. ACM, New York.

CHRISTEN, P. (2012). *Data Matching*: *Concepts and Techniques for Record Linkage*, *Entity Resolution*, *and Duplicate Detection*. Springer, Berlin.

CLOGG, C. C. (1995). Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (G. Arminger, C. C. Clogg and S. ME, eds.) 311–359. Plenum, New York.

COLLINS, J. and HUYNH, M. (2014). Estimation of diagnostic test accuracy without full verification: A review of latent class methods. *Stat*. *Med*. **33** 4141–4169. MR3267401

COOK, L. J., OLSON, L. M. and DEAN, J. M. (2001). Probabilistic record linkage: Relationships between file sizes, identifiers and match weights. *Methods Inf*. *Med*. **40** 196–203.

DAGGY, J., XU, H., HUI, S. L., GAMACHE, R. E. and GRANNIS, S. J. (2013). A practical approach for incorporating dependence among fields in probabilistic record linkage. *BMC Med*. *Inform*. *Decis*. *Mak*. **13** 97.

DAGGY, J., XU, H., HUI, S. and GRANNIS, S. (2014). Evaluating latent class models with conditional dependence in record linkage. *Stat*. *Med*. **33** 4250–4265. MR3267408

DENDUKURI, N. and JOSEPH, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* **57** 158–167. MR1833302

ESPELAND, M. A. and HANDELMAN, S. L. (1989). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* **45** 587–599.

FELLEGI, I. P. and SUNTER, A. B. (1969). A theory for record linkage. *J*. *Amer*. *Statist*. *Assoc*. **64** 1183–1210.

FORTINI, M., LISEO, B., NUCCITELLI, A. and SCANU, M. (2001). On Bayesian record linkage. *Res*. *Off*. *Stat*. **4** 185–198.

FORTINI, M., NUCCITELLI, A., LISEO, B. and SCANU, M. (2002). Modeling issues in record linkage: A Bayesian perspective. In *Proceedings of the Section on Survey Research Methods*. Amer. Statist. Assoc., Alexandria, VA.

GARRETT, E. S. and ZEGER, S. L. (2000). Latent class model diagnosis. *Biometrics* **56** 1055–1067. MR1815583

GOMATAM, S., CARTER, R., ARIET, M. and MITCHELL, G. (2002). An empirical comparison of record linkage procedures. *Stat*. *Med*. **21** 1485–1496.

GRANNIS, S. J., OVERHAGE, J. M. and MCDONALD, C. J. (2004). Real world performance of approximate string comparators for use in patient matching. *Stud*. *Health Technol*. *Inform*. **107** 43–47.

GRANNIS, S. J., OVERHAGE, J. M., HUI, S. L. and MCDONALD, C. J. (2003). Analysis of a probabilistic record linkage technique without human review. *AMIA Annu*. *Symp*. *Proc*. **2003** 259–263.

HADGU, A. and QU, Y. (1998). A biomedical application of latent class models with random effects. *J*. *R*. *Stat*. *Soc*. *Ser*. *C*. *Appl*. *Stat*. **47** 603–616.

HAGENAARS, J. A. (1988). Latent structure models with direct effects between indicatirs, local dependence models. *Sociol*. *Methods Res*. **16** 379–405.

HAN, H., GILES, L., ZHA, H., LI, C. and TSIOUTSIOULIKLIS, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the* 4*th ACM/IEEE-CS Joint Conference on Digital Libraries* 296–305. ACM, New York.

HAND, D. J. (2006). Classifier technology and the illusion of progress. *Statist*. *Sci*. **21** 1–34. MR2275965

HAND, D. and YU, K. (2001). Idiot's Bayes—Not so stupid after all? *Int*. *Stat*. *Rev*. **69** 385–398.

HERZOG, T. N., SCHEUREN, F. J. and WINKLER, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer, Berlin.

KELLY, R. P. (1986). Robustness of the Census Bureau's record linkage system. *Proc. Sect. Surv. Res. Methods* **1986** 620–624.

LARSEN, M. D. (1997). Modeling issues and the use of experience in record linkage. In *Record Linkage Techniques*—1997: *Proceedings of an International Workshop and Exposition* (W. Alvey and B. Jamerson, eds.) 95–105. National Academies Press, Washington, DC.

LARSEN, M. D. (2004). Record linkage using finite mixture models. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (A. Gelman and X.-L. Meng, eds.). *Wiley Ser. Probab. Stat.* 309–318. Wiley, Chichester. MR2138266

LARSEN, M. D. (2012). An experiment with hierarchical Bayesian record linkage. Preprint. Available at arxiv:1212.5203.

LARSEN, M. D. and RUBIN, D. B. (2001). Iterative automated record linkage using mixture models. *J. Amer. Statist. Assoc.* **96** 32–41. MR1973781

LI, X. and SHEN, C. (2013). Linkage of patient records from disparate sources. *Stat. Methods Med. Res.* **22** 31–38. MR3190644

LI, X., XU, H., SHEN, C. and GRANNIS, S. (2018). Automated linkage of patient records from disparate sources. *Stat. Methods Med. Res.* **27** 172–184. MR3745662

MARTINS, B. (2011). A supervised machine learning approach for duplicate detection over gazetteer records. In *GeoSpatial Semantics*: 4*th International Conference*, *GeoS* 2011, *Brest*, *France*, *May* 12–13, 2011. *Proceedings* (C. Claramunt, S. Levashkin and M. Bertolotto, eds.) 34–51. Springer, Berlin.

NEWCOMBE, H. B. and KENNEDY, J. M. (1962). Record linkage: Making maximum use of the discriminating power of identifying information. *Commun. ACM* **5** 563–566.

ONG, T. C., MANNINO, M. V., SCHILLING, L. M. and KAHN, M. G. (2014). Improving record linkage performance in the presence of missing linkage data. *J. Biomed. Inform.* **52** 43–54.

QU, Y., TAN, M. and KUTNER, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52** 797–810. MR1411731

QUANTIN, C., BINQUET, C., BOURGUARD, K., ALLAERT, F. A., FERDYNUS, C., PATTISINA, R., HARMENIL, G., PEGUIGNOT, S. and GOUYON, J. B. (2004). Assessment of the discriminating power of identifiers for record linkage. *Rev. épidémiol. Santé Publique* **52** 431–440.

SADINLE, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Ann. Appl. Stat.* **8** 2404–2434. MR3292503

SADINLE, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *J. Amer. Statist. Assoc.* **112** 600–612. MR3671755

SADINLE, M. and FIENBERG, S. E. (2013). A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. *J. Amer. Statist. Assoc.* **108** 385–397. MR3174628

SARIYAR, M., BORG, A. and POMMERENING, K. (2011). Controlling false match rates in record linkage using extreme value theory. *J. Biomed. Inform.* **44** 648–654.

SCHMIDTMANN, I., HAMMER, G., SARIYAR, M. and GERHOLD-AY, A. (2009). Evaluation des Krebsregisters NRW Schwerpunkt record linkage. Technical report, Institute of Medical Biostatistics, Epidemiology and Informatics at Johannes Gutenberg Univ. Available at http://www.krebsregister.nrw.de/fileadmin/user_upload/dokumente/Evaluation/EKR_NRW_Evaluation_Abschlussbericht_2009-06-11.pdf.

SEPÚLVEDA, R., VICENTE-VILLARDÓN, J. L. and GALINDO, M. P. (2008). The biplot as a diagnostic tool of local dependence in latent class models: A medical application. *Stat. Med.* **27** 1855–1869. MR2420349

SUBTIL, A., DE OLIVEIRA, M. R. and GONÇALVES, L. (2012). Conditional dependence diagnostic in the latent class model: A simulation study. *Statist. Probab. Lett.* **82** 1407–1412. MR2929794

TANCREDI, A. and LISEO, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Ann. Appl. Stat.* **5** 1553–1585. MR2849786

THIBAUDEAU, Y. (1993). The discrimination power of dependency structures in record linkage. *Surv. Methodol.* **19** 31–38.

TORRA, V., NAVARRO-ARRIBAS, G. and ABRIL, A. (2010). Supervised learning for record linkage through weighted means and OWA operators. *Control Cybernet.* **39** 1011–1026.

TORRANCE-RYNARD, V. L. and WALTER, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test preformance. *Stat. Med.* **97** 2157–2175.

TREERATPITUK, P. and GILES, C. L. (2009). Disambiguating authors in academic publications using random forests. In *Proceedings of the* 9*th ACM/IEEE-CS Joint Conference on Digital Libraries* 39–48. ACM, New York.

TROMP, M., MERAY, N., RAVELLI, A. C. J., REITSMA, J. B. and BONSEL, G. J. (2008). Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies. *J. Am. Med. Inform. Assoc.* **15** 654–660.

TROMP, M., RAVELLI, A. C. J., BONSEL, G. J., HASMAN, A. and REITSMA, J. B. (2011). Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *J. Clin. Epidemiol.* **64** 565–572.

UEBERSAX, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Appl. Psychol. Meas.* **23** 283–297.

VACEK, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* **41** 959–968.

VENTURA, S. L. and NUGENT, R. (2014). Hierarchical linkage clustering with distributions of distances for large-scale record linkage. In *Privacy in Statistical Databases*: *UNESCO Chair in Data Privacy*, *International Conference*, *PSD* 2014, *Ibiza*, *Spain*, *September* 17–19, 2014. *Proceedings* (J. Domingo-Ferrer, ed.). Springer, Berlin.

VENTURA, S. L., NUGENT, R. and FUCHS, E. R. H. (2015). Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Res. Policy* **44** 1672–1701.

WALTER, S. D., MACASKILL, P., LORD, S. J. and IRWIG, L. (2012). Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Stat. Med.* **31** 1129–1138. MR2925684

WINKLER, W. (1988). Using the EM algorithm for weight computation in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 667–671. Amer. Statist. Assoc., Providence, RI.

WINKLER, W. (1989). Methods for adjusting for lack of independence in an application of the Fellegi–Sunter model of record linkage. *Surv. Methodol.* **15** 101–117.

WINKLER, W. (1993). Improved decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 274–279. Amer. Statist. Assoc., Providence.

XU, H., BLACK, M. A. and CRAIG, B. A. (2013). Evaluating accuracy of diagnostic tests with intermediate results in the absence of a gold standard. *Stat. Med.* **32** 2571–2584. MR3067408

XU, H. and CRAIG, B. A. (2009). A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics* **65** 1145–1155. MR2756502

XU, H., LI, X., SHEN, C., HUI, S. L. and GRANNIS, S. (2019). Supplement to "Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter?" DOI:10.1214/19-AOAS1256SUPP.

YANG, I. and BECKER, B. P. (1997). Latent variable modeling of diagnostic accuracy. *Biometrics* **53** 948–958.

ZHU, R., ZHANG, J., ZHANG, D. and YAN, G. (2010). Stepwise variable selection for loglinear mixtures in record linkage. *Eur. J. Pure Appl. Math.* **3** 141–162. MR2630106

H. Xu
X. Li
Department of Biostatistics
Indiana University
Indianapolis, Indiana 46202
USA
E-mail: huipxu@iu.edu
xiaochun@iu.edu

C. Shen
Beth Israel Deaconess Medical Center
Harvard Medical School
Boston, Massachusetts 02215
USA
E-mail: cshen1@bidmc.harvard.edu

S. L. Hui
Regenstrief Institute, Inc.
1101 W 10th St.
Indianapolis, Indiana 46202
USA
E-mail: shui@iupui.edu

S. Grannis
Regenstrief Institute, Inc.
1101 W 10th St.
Indianapolis, Indiana 46202
USA
and
Department of Family Medicine
Indiana University
Indianapolis, Indiana 46202
USA
E-mail: sgrannis@regenstrief.org