Journal of Survey Statistics and Methodology (2020) ***8**, 483–512*

HOW LINKAGE ERROR AFFECTS HIDDEN MARKOV MODEL ESTIMATES: A SENSITIVITY ANALYSIS

PAULINA PANKOWSKA*
BART F.M. BAKKER
DANIEL L. OBERSKI
DIMITRIS PAVLOPOULOS

Hidden Markov models (HMMs) are increasingly used to estimate and correct for classification error in categorical, longitudinal data, without the need for a "gold standard," error-free data source. To accomplish this, HMMs require multiple observations over time on a single indicator and assume that the errors in these indicators are conditionally independent. Unfortunately, this "local independence" assumption is often unrealistic, untestable, and a source of serious bias. Linking independent data sources can solve this problem by making the local independence assumption plausible across sources, while potentially allowing for local dependence within sources. However, record linkage introduces a new problem: the records may be erroneously linked or incorrectly not linked. In this paper, we investigate the effects of linkage error on HMM estimates of transitions between employment contract types. Our data come from linking a labor force survey to administrative employer records; this linkage yields two indicators per time point that are plausibly conditionally independent. Our results indicate that both false-negative and false-positive linkage error turn out to be problematic primarily if the error is large and highly correlated with the dependent variable. Moreover, under certain conditions, false-positive linkage error (mislinkage) in fact

Paulina Pankowska, Bart F.M. Bakker, and Dimitris Pavlopoulos are with the Vrije Universiteit Amsterdam. Bart Bakker is with the Statistics Netherlands. Daniel L. Oberski is with the Universiteit Utrecht. This work was supported by Statistics Netherlands and Vrije Universiteit Amsterdam; the authors would like to thank Peter Paul de Wolf (Statistics Netherlands) and the members of the SILC research group of the Vrije Universiteit Amsterdam for reviewing the paper, as well as Richard Price for both reviewing and editing the paper and Aleksandra Daniłoś for assisting with the design of the figures. This paper is a part of a PhD project co-funded by Statistics Netherlands (in Dutch: Centraal Bureau voor de Statistiek, CBS).

*Address correspondence to Paulina Pankowska; Department of Sociology, Hoofd Gebouw (HG), De Boelelaan 1105, 1081HV Amsterdam, the Netherlands. E-mail: p.k.p.pankowska@vu.nl.

doi: 10.1093/jssam/smz011 Advance access publication 29 May 2019

acts as another source of misclassification that the HMM can absorb into its error-rate estimates, leaving the latent transition estimates unbiased. In these cases, measurement error modeling already accounts for linkage error. Our results also indicate where these conditions break down and more complex methods would be needed.

KEYWORDS: Classification error; Hidden Markov model (HMM); Linkage error; Measurement error; Misclassification; Record linkage.

1. INTRODUCTION

Despite numerous efforts to the contrary, survey and register data almost inevitably contain measurement error (Kuha and Skinner 1997; Biemer and Stokes 2004; Alwin 2007). Such errors severely bias estimates of relationships between variables and, therefore, it is essential to account and correct for them (Fuller 1987; Kuha and Skinner 1997; Carroll, Ruppert, Stefanski, and Crainiceanu 2006; Saris and Gallhofer 2007). For categorical variables, an attractive method of doing so—without requiring "gold standard" (error-free) validation data—is latent class models (LCMs) (Vermunt and Magidson 2002; Biemer 2011).

The LCMs use repeated indicators of some categorical phenomenon of interest as input, and output estimates of the classification error rates of these indicators, otherwise known as "measurement parameters." These models also provide estimates of the "structural parameters," which measure quantities of scientific interest, such as prevalence of certain groups in the population or transitions over time. If the repeated indicators—which are used as inputs as part of a set of different survey questions or different administrative records—are intended to measure a single underlying latent variable, the LCM becomes a "latent structure model." When the repeated indicators are repetitions of the same question or administrative record at different time points, a particular variant of an LCM is used: the "hidden" (or "latent") Markov model (HMM) (Alwin 2007; Alwin, Baumgartner, and Beattie 2017). In this paper, we focus on HMMs, which are regularly applied to categorical longitudinal data (Biemer 2011; Biemer, De Leeuw, Eckman, Edwards, Kreuter, et al. 2017; Edwards, Berzofsky, and Biemer 2017).

The great advantage of LCMs is that all indicators are allowed to contain errors and, as such, LCMs can estimate the quality of a survey indicator without requiring perfect comparison data. However, this exciting feature of LCMs does not come cheap: a payment in untestable assumptions is required, in particular the "local independence" assumption, which demands that the errors in the repeated indicators occur independently (see, e.g., Oberski, Hagenaars, and Saris 2015).

This local independence assumption is unrealistic, harmful and, when only one indicator is available, also undetectable. It is *unrealistic*, because "common method variance"—that is, variance attributed to the measurement method as opposed to the constructs the measure represents—is typically found in studies able to detect it (Saris and Gallhofer 2007), and because it is likely that any personal "style" in answering a survey question carries over time (Billiet and Davidov 2008). It is also highly probable that specific errors in registers are repeated for a certain period of time (Pavlopoulos and Vermunt 2015). It is *harmful* because ignoring it leads to bias in the HMM parameter estimates (Vacek 1985; Torrance-Rynard and Walter 1997; Georgiadis, Johnson, Gardner, and Singh 2003; Qu and Hagdu 2012); Appendix A.1 provides an illustration of the severity of the bias using employment mobility data from the Netherlands. Finally, it is *undetectable* with data from a single repeated indicator because the local independence assumption is necessary for model identification in this case. While it is, in general, possible to detect and model local dependence in LCMs (Hagenaars 1988; Oberski 2016), in HMMs, the parameters that represent local dependence are only generally identifiable if a second indicator of the variable of interest is obtained at each time point (Hagenaars 1990). Such an indicator should then plausibly contain errors that are independent of the errors present in the first indicator.

Therefore, an attractive solution to the problem of local independence is to link different data sources, such as surveys and administrative registers. The attractiveness of this solution lies in the fact that neither of these two data sources is required to be error-free; it is only required that the survey errors are independent of the register errors, which indeed seems plausible. This means that, by combining registers and surveys, it becomes possible to allow for local dependence within each source. Previous studies have done so, and indeed have found considerable local dependence (Bassi, Hagenaars, Croon, and Vermunt 2000; Pavlopoulos and Vermunt 2015; Oberski, Kirchner, Eckman, and Kreuter 2017), confirming both the importance of relaxing this assumption and the attractiveness of data linkage.

Record linkage allows us to tackle the problems of measurement error modeling, but it introduces a new challenge: linkage error. Such errors, which occur when records of different individuals are wrongly linked or when records of the same individuals are wrongly not linked, are known to bias estimates of interest when left unaccounted for (Harron, Doidge, Knight, Gilbert, Goldstein, et al. 2017). Several estimators correcting for linkage errors have been suggested (e.g., Lahiri and Larsen 2005; Chambers 2009; Liseo and Tancredi 2011; Goldstein, Harron, and Wade 2012); Di Consiglio and Tuoto (2018) show that these methods are effective in reducing linkage error bias in linear and logistic regression analyses. However, some of these estimators assume knowledge of the posterior probability of correct linkage for all pairs of cases. This knowledge is unavailable to most analysts in practice. The remaining solutions do not assume this knowledge, but have only been developed for regression models (Chambers 2009).

In this paper, we study the extent to which linkage error biases HMM parameter estimates. Through a simulation study based on a real data application to linked survey-register employment records at Statistics Netherlands, we demonstrate the sensitivity of the structural (transition rate) parameters of the model to linkage error. We find that in certain situations, the HMM can absorb the error into its measurement model, leading to approximately unbiased structural parameter estimates. In other situations, however, non-negligible biases in the structural part of the model do occur. A novel geometric representation of the latent class estimation problem demonstrates why this is the case.

Section 2 first provides some background information on single- and multiple-indicator hidden Markov models and then discusses the topic of linkage error and its effects on HMMs. Section 3 presents the data and section 4 the methodology; in section 5 we discuss the results of our analysis. Section 6 provides conclusions.

2. BACKGROUND

2.1 Hidden Markov Models (HMMs) and Measurement Error

Hidden Markov models (HMMs) are a group of latent class models that are increasingly used to estimate and correct for measurement error in longitudinal categorical data (Biemer 2004, 2011). In this section, we first present the basic single-indicator HMM, commonly applied across the literature; we then extend it by including an additional indicator per time point.

The basic HMM operates under the assumption that, at each time point $t \in \{1, \ldots, T\}$, the observed answer Y_t is assumed to follow a multinomial distribution and is generated *independently* with some probability $P(Y_t|X_t)$ from the true, but unobserved, multinomially distributed variable X_t. Because the generation of Y_t is assumed independent of all other variables, the T-dimensional distribution $P(Y|X)$ of observed path Y given latent path X, where $X = (X_1, \ \ldots, \ X_T)$, factorizes into the following product:

$$P(Y|X) = \prod_{t=1}^{T} P(Y_t|X_t)\,. \tag{1}$$

This assumption is known as the "local independence" or "independent classification error" (ICE) assumption. The latent path X, meanwhile, is assumed to follow a Markov or an *AR(1)* process,

$$P(X) = \ P(X_0)\prod_{t=1}^{T} P(X_t|X_{t-1}). \tag{2}$$

Finally, the observed data distribution $P(Y)$ is assumed to arise by combining the ICE and Markov assumptions that are mentioned above and then marginalizing over X. This yields the following marginal likelihood:

$$P_{\text{HMM}}(Y) = \sum_X P(Y|X)P(X), \tag{3}$$

with "structural" parameters $P(X_0)$ and $P(X_T|X_{T-1})$—which correspond to the initial state and transition probabilities—and "measurement parameters" $P(Y_t|X_t)$—which are the probabilities of correct and incorrect classification.

When consistent estimates of $P_{\text{HMM}}(Y)$ are observed (i.e., when Y is "ergodic"), consistent maximum-likelihood estimates can be obtained by maximizing (3) over the structural and measurement parameters (Leroux 1992). In practice, instead of the exponentially complex summation over all possible latent paths X in (3), the more computationally efficient "forward-backward" (Baum-Welch) algorithm is used. This amounts to an adapted expectation-maximization (EM) procedure (McLachlan and Krishnan 2008, pp. 291–2). In the E-step of this procedure, the posterior probability $P(X|Y)$ is estimated by combining two computational steps—the forward and backward recursions. Specifically, in the forward step, the algorithm calculates the probability of arriving at a specific state at time t given the states that occurred up until that time point; in the backward step, this probability is calculated based on the states occurring at time points following t. Thus, each of the steps considers one time point at a time but in combination with the results of the respective previous computations. In the M-step, the model's parameters are computed by summing over the states at each time point. This sum is weighted by the posterior probabilities. Thus, the computational complexity of one Baum-Welch iteration is linear in the number of time points, rather than exponential, as when using the marginal likelihood (3). The E- and M-steps are iterated until convergence is reached.

The single-indicator HMM is attractive for two reasons. First, in contrast with standard latent class analysis, it allows for hidden change over time in the true values, $P(X_t|X_{t-1})$, while simultaneously estimating and accounting for classification errors, $P(Y_t \neq x_t|X_t = x_t)$. Second, its parameters can be identified from panel data on single repeated indicators with three or more waves, which are often already collected as part of longitudinal surveys or recorded in administrative databases. This identifiability follows from the model's assumptions, specifically the Markov and conditional independence (ICE) assumptions.

However, as already discussed in the introduction, conditional independence may in practice be an unrealistic assumption. To model such error dependencies and simultaneously estimate classification error in both survey and administrative data that measure the same phenomena, Pavlopoulos and Vermunt (2015) suggest linking respondents' survey answers to administrative records.

Such linked survey-administrative data then allow for the relaxation of the ICE assumption, replacing (1) with

$$P(Y|X) = P\left(Y_{\text{survey}}|X\right)P(Y_{\text{admin}}|X), \tag{4}$$

where Y now collects the observed processes for both survey and administrative data. Pavlopoulos and Vermunt (2015) suggest further specifying the conditional dependence as

$$P(Y|X) = \prod_{t=1}^{T} P\left(Y_{t,\text{survey}}|X_t\right) \prod_{t=1}^{T} P\left(Y_{t,\text{admin}}|X_t, X_{t-1}, Y_{t-1,\text{admin}}\right), \tag{5}$$

with $P\left(Y_{t,\text{admin}}|X_t, X_{t-1}, Y_{t-1,\text{admin}}\right)$ modeled by logistic regression. This model allows for error dependence in the administrative data, while assuming survey and administrative answers to be conditionally independent. The advantages of record linkage are thus that (1) both survey and administrative errors can be modeled simultaneously, and (2) the ICE assumption can be relaxed in a rather flexible way.

The disadvantage of linkage, however, is that linkage error may occur and cause bias in analyses of dependencies, such as linear and logistic regression (Chambers and Kim 2016). Therefore, it seems plausible that bias would also occur in a multivariate method such as HMM, which uses dependencies to estimate its parameters. However, no work to date has examined the precise effects of linkage error for this specific group of models. This paper does not aim to examine these effects analytically or solve the problem of linkage error for HMMs. We do, however, note that linkage error can be expected to strongly violate HMM assumptions and cause bias under certain circumstances. In the next section, we provide an intuitive explanation of this phenomenon. In doing so, we first provide a formal definition of record linkage and the errors associated with it; we then present a theoretical consideration of how linkage errors (might) affect HMMs.

2.2 Linkage, Its Associated Errors, and Their Effects on HMMs

Record linkage is a process that matches records and attempts to select those matches that belong to the same person or unit. The process uses one or more data fields (i.e., linkage variables) that contain the same identifying information in all data sources (Armstrong and Mayda 1993; Winkler 1999).

There are two main types of record linkage methods—deterministic and probabilistic. Deterministic record linkage defines pairs as true matches if the matching variables agree exactly in all data sources. It usually relies on a relatively small number of matching variables and is most commonly applied in the presence of the same unique identifier in all data sources (Blakely and Salmond 2002). As data sources have been increasingly lacking high-quality unique identifiers, deterministic linkage has been gradually replaced by probabilistic linkage (Ariel, Bakker, de Groot, van Grootheest, van der Laan, et al. 2014).

Probabilistic record linkage tends to use a larger number of matching variables and does not require an exact agreement on all of them for a pair to be considered a true match. Probabilistic linkage determines the probability of a match being correct and, as such, whether it should be regarded as a "true" or "false" match (Fellegi and Sunter 1969; Armstrong and Mayda 1993; Blakely and Salmond 2002; Winglee, Valliant, and Scheuren 2005; Bohensky, Jolley, Sundararajan, Evans, Pilcher, et al. 2010).

While record linkage is undoubtedly an important tool that allows combining information from various sources, it is also associated with different types of errors. In general, linkage errors occur: (1) when due to missing or inaccurate data, some records that correspond to the same person or unit are not linked—a phenomenon referred to a *false-negative* linkage error—and (2) when as a result of coding or measurement errors, unrelated records are wrongfully linked—a situation referred to as a *false-positive* linkage error (Winglee et al. 2005; Bohensky et al. 2010).

Record linkage and linkage errors can be formulated using files drawn from two populations—file A containing N_A records and file B containing N_B records, and a set C containing record pairs that are the cross-product of files A and B. This set is denoted by $C = \{(a, b); a \in A,\ \ b \in B\}$, and the number of records equals $N = \ N_A \ \times \ N_B$ (Armstrong and Mayda 1993; Sadinle, Hall, and Fienberg 2011).

The aim of record linkage is to divide set C into two separate sets— one that includes true matches (here denoted by M) and one that includes true non-matches (here denoted by U). This is often done by examining the data contained in files A and B and deciding whether the records certainly belong to the same entity (i.e., are a definite link, denoted by A_1), possibly belong to the same entity (i.e., are a possible link, denoted by A_2), or certainly belong to different entities (i.e., are a definite non-link, denoted by A_3) (Fellegi and Sunter 1969; Armstrong and Mayda 1993; Sadinle et al. 2011).

False-positive and false-negative types of error occur respectively when (1) a record pair that belongs to the true non-match set (U) is registered as a link (A_1) and (2) when a record pair belonging to the true match set (M) is registered as a non-link (A_3). Thus, the false-positive linkage error can be denoted by $P(A_1|U)$ and the false-negative by $P(A_3|M)$ (Armstrong and Mayda 1993; Sadinle et al. 2011).

There are several approaches and frameworks available in the literature to correct for the effects of linkage error. Three prominent approaches are those proposed by Lahiri and Larsen (2005), Chambers (2009), and Liseo and Tancredi (2011). Lahiri and Larsen (2005) propose an M- and U- probabilities-weighted linear regression model for linked data, which takes into account linkage uncertainty. However, their method relies on the assumption that the linkage/mislinkage probabilities of all pairs of records are known to the analyst. This assumption is often unrealistic in practice. Liseo and Tancredi (2011) propose a Bayesian approach to linkage problems, in which the analysis and linkage models are subsumed into a single latent variable model estimated via Markov Chain Monte Carlo (MCMC). A similar approach, implementing Bayesian imputation conditioned on the linkage probabilities, is suggested independently by Goldstein et al. (2012). Other studies that propose Bayesian approaches to correct for linkage error include those by Sadinle (2014, 2017), Steorts (2015), and Steorts, Hall, and Fienberg (2016). While the Bayesian approach is, in principle, comprehensive, it shares the drawback of the approach of Lahiri and Larsen (2005) that full knowledge of the linkage process is required by the analyst. Finally, Chambers (2009) and Kim and Chambers (2012a, 2012b) introduce a bias-corrected ratio estimator, as well as a class of weighted estimators for linear regression and logistic regression. Moreover, Chambers (2009) suggests replacing the required assumption of perfect information regarding the linkage/mislinkage probabilities with a more realistic approximation based on available aggregate linkage rates. As detailed in Chambers and Kim (2016), since the weighting approach is based on estimating equations, it can in principle be extended to other, more complex, classes of models beyond linear and logistic regression. However, Chambers-type estimators for HMMs are currently not available.

To sum up, the available methods to account for linkage error are difficult to implement for HMMs for practical or technical reasons. It is therefore important to investigate the sensitivity of such models to linkage error, which is the focus of this paper. While false-negative linkage error manifests itself as missing data, and a large literature on the effects of various missingness mechanisms on maximum-likelihood (ML) estimates already exists (see, e.g., Little and Rubin 2002), false-positive linkage errors (mislinkages) have an entirely different, as yet unstudied, effect on HMMs. Therefore, our theoretical considerations elaborate on the effect of mislinkages; the simulation study, however, investigates the effects of both types of linkage errors.

Following Lahiri and Larsen (2005), mislinkage among declared links manifests itself as an additional latent class variable with two categories corresponding to true matches (M) and non-matches (U). Within the class of matches, the HMM holds, while within the class of non-matches an unknown process holds. Lahiri and Larsen (2005) assume non-matches to follow a distribution in which all J observed variables are independent. The observed data distribution is then a mixture of the true dependence structure and “randomly shuffled” data:

$$P_{\text{linked}}(Y) = P(M)P_{\text{HMM}}(Y|\theta) + [1 - P(M)]\prod_{j=1}^{J} P(Y_j), \tag{6}$$

where the HMM likelihood has been expressed as $P_{\text{HMM}}(Y|\theta)$ to emphasize its dependence on the model parameters of interest, θ. Clearly, when fitting the

HMM to P_{linked}, asymptotic bias may, in principle, occur whenever there is mislinkage. Intuitively, however, unless the mixture P_{linked} induces additional dependence beyond that found in P_{HMM}, its effect is to increase random measurement error in each Y_j. Since the HMM is intended to capture such errors and correct for them, one might expect that the increased error rates are reflected in the measurement part of the model which describes $P(Y|X)$ but not necessarily in the structural model describing $P(X)$. Appendix A.2 argues geometrically that, when the linkage error is independent of Y, this intuition will hold approximately. In particular, we show that the maximum likelihood solution for the "structural" parameter indicating the class size, π, is approximately unaffected by independent linkage error. In the following sections, a simulation study investigates the extent to which this result holds in an HMM.

3. DATA

The data in our analysis are from the Dutch Labour Force Survey (LFS) and the Employment Register (ER), which have been linked using each citizen's unique identification number or the combination of birth date, sex, postal code, and house number as a linkage key. In this paper, we assume that this process does not involve linkage error and simulate the effect of linkage error by artificially introducing false-negative and false-positive linkages into the dataset.

Our sample consists of 15 months of observations on 8,886 LFS respondents aged 25 to 55 who first participated in the survey in 2009. This results in a total sample size of 133,290 observations. The employment register is observed on a monthly basis, while the LFS is taken every three months and consists of five waves. The main variable of interest in our analysis is an individual's employment contract type for their primary job, which can take one of the following values: "permanent contract," "temporary contract," or "other." For further details about the dataset, see Appendix A.3.

4. METHODOLOGY

4.1 Model

Our approach consists of a simulation analysis in which we make use of a two-indicator HMM, where one of the indicators is the individual's contract type according to the LFS and the second is the contract type according to the ER. While the model could be extended further, following Pavlopoulos and Vermunt (2015) and Pankowska, Bakker, Oberski, and Pavlopoulos (2018), our simulations are based on a simplified model that retains the local independence assumption. The following equation estimates the probability of following a certain observed path according to our model:

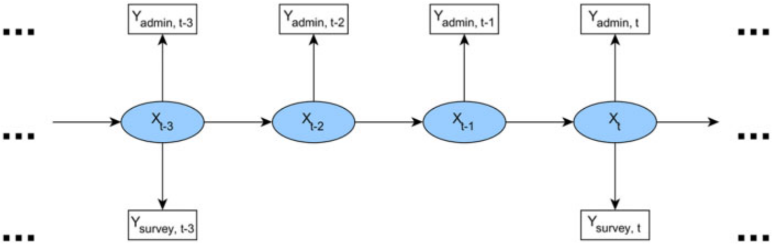

Figure 1. Hidden Markov Model Graph. Rectangles are observed variables, while ovals are latent "true" variables. Absence of arrows indicates conditional independence.

$$P\left(C_i = c_{i,} \, E_i = e_i\right) = \sum_{x_0=1}^{3} \sum_{x_1=1}^{3} \dots \sum_{x_T=1}^{3} P(X_{i0} = x_0) \prod_{t=1}^{T} P\left(X_{it} = x_t | X_{i(t-1)} = x_{t-1}\right)$$

$$\times \prod_{t=1}^{T} P(C_{it} = c_t | X_{it} = x_t) \prod_{t=1}^{T} P(E_{it} = e_t | X_{it} = x_t)^{\delta_{it}}, \quad (7)$$

where C_{it} and E_{it} denote the contract type of person i at month t according to the ER and LFS, respectively, with $i = 1, \dots,$ N and $t = 1, \dots, 17$.[1] To account for the fact that the contract type according to the survey (E_{it}) can only be observed every third month, the indicator δ_{it} is included in the model; δ_{it} equals 1 if the survey information is available for a given month and 0 if it is missing. This amounts to assuming an ignorable (MAR) missingness mechanism (Little and Rubin 2002, pp. 118–9). The model also includes a latent (unobserved) variable (X_{it}), which represents the individual's actual contract type at time t. Both the observed indicators and the latent variable (which are referred to in the model as c_t, e_t, and x_t, respectively) consist of three categories—permanent, temporary, and other type of contract.

Figure 1 illustrates our model as a graph. Because the survey has been administered once every quarter, while monthly measures are available from the administrative database, the survey is missing at timepoints *t – 1* and *t – 2*. Estimation of the latent class model with missing data proceeds using maximum likelihood under the ignorability assumption (Little and Rubin 2002, pp. 118–9; Vermunt and Magidson 2013). Standard errors of the parameters can be obtained by inverting the expected or observed information matrix of the observed-data likelihood above.

We apply the model to different conditions in which various types of either false-negative or false-positive linkage errors are introduced into the original dataset. A summary of the simulation setup is provided as a tree graph in figure 2.

[1] In our analysis, we use data from January 2009 until March 2010, which corresponds to 17 months and, therefore, *t* runs from 1 to 17.

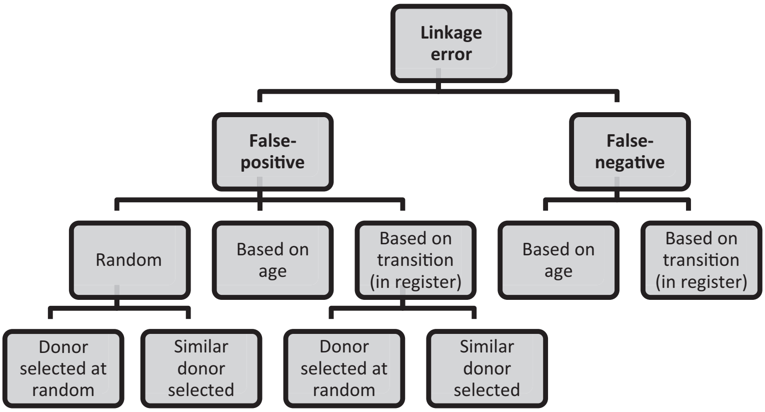

Figure 2. Conditions of the Simulation Study.

We consider conditions in which individuals are either randomly selected to be mislinked and/or excluded versus conditions in which the probabilities of linkage error depend on covariates mildly or strongly correlated with the model estimates. We also consider different error rates. Our setup allows for the investigation of the biasing effects of the error under varying degrees of severity. Each condition is replicated 200 times. We investigate the bias introduced by the error by comparing the obtained transition rates from temporary to permanent employment to the transition rates estimated using the original linked dataset. To simulate linkage error, we use the R version 3.2.3. The HMM is estimated using Latent GOLD version 4.5. For our code, please see the supplementary data online.

4.2 False-Negative Error Simulations

When investigating the effect of false-negative linkage error on the accuracy of our model estimates, we consider two conditions in which the individuals' probabilities of exclusion are correlated with (1) age[2] and (2) the presence of a (three-monthly) transition from temporary to permanent employment in the register data.[3] A condition in which the missingness is MCAR (missing completely at random) has been omitted. Within each condition, we simulate

[2] Pankowska et al. (2018) in their analysis of the same data used an extended version of the HMM we use in this paper. Their model, among other things, accounted for the effect of age on the latent transition probabilities. Their results showed that age has a moderate, negative effect on the probability of transitioning from temporary to permanent employment (logit coefficient = −0.3 over the range of the covariate).

[3] According to our model, over 99 percent of all contracts observed in ER are correctly classified and, therefore, the transition covariate we have created and the model estimates are highly correlated.

three subconditions in which we introduce high (20 percent), medium (10 percent), and low (5 percent) overall exclusion error into our data; this error is equal to the proportion of correctly linked individuals in the data that are erroneously excluded.

For the age-dependent conditions, the correlations are such that the exclusion probabilities of younger individuals are higher than those of older individuals; for the transition-dependent conditions, the probability of exclusion for those individuals who transitioned according to the register data is higher than that for the individuals who did not. These specifications are motivated by the fact that both young individuals and those who transitioned would tend to have higher residential and employment mobility and are thus more susceptible to linkage error.

To ensure that the conditions indeed represent varying levels of severity, the simulation is also designed in such a way that, as we move from conditions with lower levels of exclusion error to conditions with higher ones, the oversampling of young individuals or those who transitioned becomes more extreme (i.e., their individual exclusion probabilities increase). To illustrate, the exclusion probability of young individuals (aged 25 to 34) is set to 0.15, 0.30, and 0.70 when the overall exclusion rate is low (5 percent), medium (10 percent), and high (20 percent), respectively; the exclusion probability of older individuals (aged 35 to 54) remains at 0.01 in all three cases.

Thus, a higher level of false-negative linkage error not only indicates that a larger proportion of individuals is excluded from the sample, but it also implies that the remaining sample is less representative of the overall population in terms of characteristics that are correlated with the transition rates estimated by the model. As those covariates are not controlled for when estimating the HMM, these simulated datasets are equivalent to a dataset containing data missing not at random (MNAR).

Overall, the simulations consist of three steps. First, the exclusion rate and the individual exclusion probabilities are set; then individuals are excluded from the sample with a probability equal to that condition's exclusion probability. Finally, the HMM is fitted to the resultant subsample and the estimates are compared to those obtained when using the full sample. As an illustration, Appendix A.4.1 provides pseudocode for generating one condition.

4.3 False-Positive Error Simulations

The analysis of the false-positive linkage error, similarly to that for the false-negative, also follows three steps. Note that here, unlike in the false-negative example (whereby individuals are merely excluded from the sample), a proportion of the sample is mislinked with another set of individuals. This adds a further complication to the simulation design, as a donor is required whose ER contract type can be (erroneously) linked to a given individual's LFS contract information. As in the false-negative error conditions, the first step determines the overall level of mislinkage (5 percent, 10 percent, or 20 percent) and the individual probabilities of an erroneous link (which are either assigned at random or are age- or transition-dependent).

In the second step, the false-positive error is simulated in the following way: a number of individuals is selected at random according to the aforementioned design. Each one of those individuals in turn, here referred to as individual A, is either (1) randomly matched to another person or (2) matched to a similar person based on age, gender, education level, and ethnicity. The register values of individual A for the contract type are replaced with those of the matched individual (i.e., the donor), here referred to as individual B.

The second set of conditions, wherein relatively similar individuals are matched, is introduced to approximate a more realistic linkage error condition that is more representative of actual potential mismatches.

The third and final step is parallel to that of the exclusion error analysis. Our HMM is fitted to each of the simulated datasets, and the outcomes are compared to the results obtained when using the original dataset. Pseudocode illustrating the simulation setup for one of the conditions is included in Appendix A.4.2.

5. RESULTS

5.1 The Effect of False-Negative Error

The simulation results obtained for the various false-negative error conditions are shown in table 1; the table provides the mean estimated three-monthly transition rates as well as the absolute and relative bias introduced by linkage error. These biases are estimated by comparing the obtained transition rates to those calculated using the original dataset. Figure A.2, which is included in Appendix A.5, provides an illustration of the relationship between the type (age or transition dependent), level (5 percent, 10 percent, 20 percent), and bias introduced by linkage error.

The results show that when the exclusion probability depends on age, the relative bias introduced by false-negative linkage error does not exceed 5 percent and, therefore, can be considered negligible. Thus, it appears that when the exclusion probability depends on a covariate that is weakly or moderately correlated with the model estimates, the bias in the model estimates is marginal, even when the overall exclusion rate is rather high (e.g., 20 percent).

A vastly different picture emerges when the exclusion probability depends on whether a transition occurred. Namely, our results show that the employment transition rates in this set of conditions are heavily underestimated, leading to a substantial, non-negligible bias. In relative terms, the bias ranges from 10.6 percent, for an overall linkage error of 5 percent, to 25 percent, when the

Table 1. Simulation Results—the Biasing Effects of All False-Negative Linkage Error Conditions (in %)

Error type	Condition: the probability of being excluded	Overall error (approx.)	High exclusion probability	Low exclusion probability	Temporary to permanent transition rate		
					Transition rate	Absolute bias	Relative bias
No error	Original HMM	0	-	-	6.9	-	-
False-negative	Depends on age	5	15	1	6.6	0.3	4.6
		10	30	1	6.7	0.2	3.2
		20	70	1	6.6	0.3	3.8
	Depends on transition	5	15	5	6.2	0.7	10.6
		10	34	9	5.2	1.7	25.0
		20	90	17	1.1	5.8	84.3

In the age-dependent conditions, high exclusion probability was set for young individuals and low for older ones; in the transition-dependent conditions, high exclusion probability was set for individuals who had a transition and low for those who did not. The transition rates are estimated based on the modal class memberships (i.e., at each time point individuals are assigned the contract type to which they have the highest posterior probability of belonging according to the model); as the entropy R2 is above 0.99 for all conditions, such an assignment is not expected to produce different results from an assignment that takes the uncertainty of class memberships into account.

linkage error amounts to 10 percent, and to as high as 84.3 percent when the error rate equals 20 percent. As this covariate is highly correlated with the model estimates, we can infer from these results that conditions characterized by substantial dependency between the error and model outcomes will result in non-negligible bias.

Overall, the results obtained suggest that the extended, two-indicator HHM is robust to false-negative linkage error when the exclusion probability depends on age, a covariate that is weakly or moderately correlated with the (structural) model estimates. In these situations, the bias introduced by linkage error is relatively small and thus the HMM estimates can be considered accurate. The model appears sensitive, though, to false-negative linkage error when the individual-level exclusion probabilities depend on whether a transition occurred, a covariate that is highly correlated with the latent variable and consequently the model outcomes. These scenarios lead to a substantial, non-ignorable bias.

Finally, it is worthwhile to note that our false-negative linkage error analysis can be viewed as a form of complete case analysis with varying degrees of missingness. Our two specific sets of conditions mimic MNAR: first, where the exclusion probabilities are dependent on a variable that is moderately correlated with the model estimates; and second, where the probabilities are dependent on a variable exceptionally highly correlated with the model estimates. Our findings confirm this line of thought. More specifically, our results, similar to those reported by studies investigating missingness specifically, show that MNAR leads to substantial bias when the missingness is highly correlated with model estimates (Marshall, Altman, Royston, and Holder 2010; Bakker and Daas 2012; Galimard, Chevret, Protopopescu, and Resche-Rigon 2016).

5.2 The Effect of False-Positive Error

The results obtained when simulating various levels and types of false-positive linkage error are presented in table 2 and in figure A.3, which is included in Appendix A.5. As can be seen, the bias introduced by false-positive linkage error is rather modest for the conditions where the mislinkage probability is either random or depends on age. In contrast, those conditions in which the probability of mislinkage depends on whether a transition occurred are characterized by high, non-negligible bias. These findings are consistent for both the conditions in which an individual is mislinked with a randomly selected donor and where the individual is mislinked with a donor similar to them with regard to age, gender, education, and ethnicity. While in the present case mislinking similar donors did not reduce the linkage error bias, this will not necessarily always be the case. If both the mislinkage probability and donor matching depend on a variable(s) that is (are) highly correlated with the transition estimates, it is likely that using similar rather than random donors would decrease the bias introduced by linkage error.

Table 2. Simulation Results—the Biasing Effects of All False-Positive Linkage Error Conditions (in %)

Error type	Condition: the probability of being mislinked	Overall error (approx.)	High exclusion probability	Low exclusion probability	Temporary to permanent transition rate		
					Transition rate	Absolute bias	Relative bias
No error	Original HMM	0	-	-	6.9	-	-
False-positive; mislinkage with random donor	Random	5	-	-	6.9	0.0	0.1
		10	-	-	6.9	0.0	0.3
		20	-	-	6.8	0.1	1.0
	Depends on age	5	15	1	6.9	0.0	0.3
		10	30	1	6.8	0.1	1.2
		20	70	1	6.7	0.2	2.6
	Depends on transition	5	15	5	6.4	0.5	7.8
		10	34	9	5.5	1.4	20.7
		20	90	17	2.4	4.5	64.6
False-positive; mislinkage with similar donor	Random	5	-	-	6.7	0.2	3.2
		10	-	-	6.7	0.2	3.2
		20	-	-	6.6	0.3	4.9
	Depends on transition	5	15	5	6.1	0.8	11.5
		10	34	9	5.1	1.8	26.6
		20	90	17	1.2	5.7	82.6

In the age-dependent conditions, high exclusion probability was set for young individuals and low for older ones; in the transition-dependent conditions, high exclusion probability was set for individuals who had a transition and low for those who did not. The results from the random and age-based mislinkage, when individuals are mislinked with random donors, were very similar, and therefore when individuals were mislinked with similar donors, the age-based set of conditions was omitted. The differences in the bias obtained when using random and similar donors might be due to the fact that the StatMatch R package used to match donors does not allow for missing values on the covariates and, thus, the analysis was run on a smaller sample. The transition rates are estimated based on the modal class memberships (i.e., at each time point individuals are assigned the contract type to which they have the highest posterior probability of belonging according to the model); as the entropy R2 is above 0.99 for all conditions, such an assignment is not expected to produce different results from an assignment that takes the uncertainty of class memberships into account.

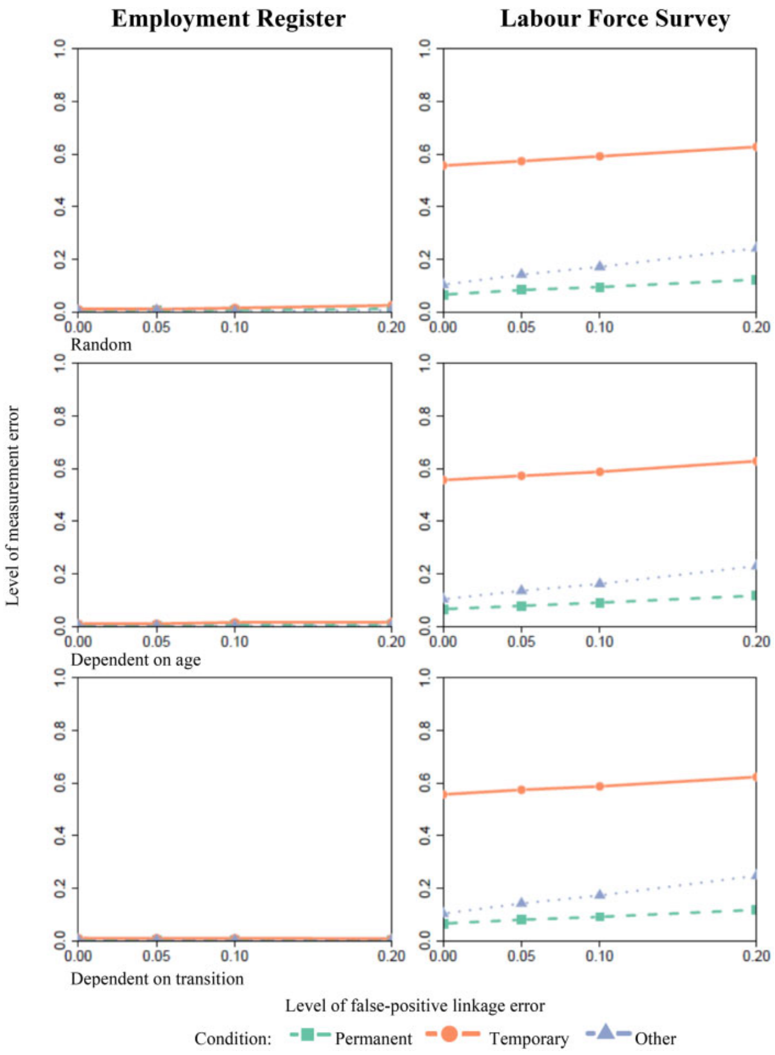

Figure 3. Level of Measurement Error by Type and Level of Mislinkage.

More specifically, the first two sets of conditions, regardless of whether the individual is mislinked with a random or a similar donor, lead to a relative bias of less than 5 percent. On the other hand, those conditions in which the mislinkage probability depends on the presence of a transition result in a relative bias of around 10 percent, 20–25 percent, and (well) over 60 percent when the mislinkage rate is low, medium, and high, respectively. Figure A.3 shows a clear positive relationship between the transition-dependent mislinkage level and the bias in the model estimates. This relationship is not observed for the other two sets of conditions.

Figure 3 demonstrates how mislinkage affects the measurement part of the model; that is, it shows the effect of linkage error on the proportion of measurement error in our main variable of interest, i.e., the individual's contract type. As can be seen, as we increase the mislinkage rate, the misclassification rate moves in tandem; this is particularly visible for the LFS data.[4] These results confirm our intuition and suggest that under many conditions false-positive linkage error is simply another source of misclassification that the HMM can absorb into the error rate estimates and correct for in the transition rate estimates.

6. CONCLUSION AND DISCUSSION

Latent class models (LCMs) have been increasingly used to correct for measurement error in categorical variables. A particularly useful group of LCMs are hidden Markov models (HMMs), as they can be applied to longitudinal data, and thus allow the study of transitions and change over time, which is often a quantity of interest in the social sciences. However, while HMMs are an appealing and useful tool, they rely on the (often unrealistic) local independence assumption. An attractive solution that allows the local independence assumption to be relaxed is linking data from independent sources. Such record linkage identifies HMMs with local dependence within sources while maintaining the independence assumption across sources. However, this approach introduces a new challenge: linkage error.

In this paper, we investigate the sensitivity of HMM estimates to linkage error. A geometric argument demonstrated that independent (false-positive) linkage error is largely absorbed by measurement parameters of latent class models. Dependent linkage errors, however, can be expected to strongly bias structural model parameters such as the latent class size in an LCM. Our simulation study further investigated this effect for HMMs based on an existing application to linked data on employment mobility.

Our results suggest that linkage error may not always be a problem for researchers who wish to apply HMMs for the purpose of estimating their structural parameters, such as transition rates. When individuals are randomly mislinked or not linked, the resulting bias in structural parameters was often negligible in our study, a result that confirms the geometric intuition relevant to LCMs. Linkage error led to significant bias only when the individual

[4] This pattern is not observed in the ER data, as the simplified HMM we use does not account for autocorrelation of the error in these data. As measurement error in the ER is predominantly systematic, the model fails to capture it altogether and assumes the register data to be virtually error-free.

probability of being erroneously excluded or mislinked depended on the transition rate itself. The bias was particularly high for high rates of linkage error and when the aforementioned dependency was very strong; in the other instances investigated, the sensitivity of estimates of structural parameters to mislinkage appears relatively low.

Our results show that false-positive linkage error can often be absorbed by the model. In other words, mislinkage can manifest itself as random measurement error that is already corrected for by the model, unless the linkage error probability is strongly dependent. Despite this important caveat, we believe that our findings highlight the attractiveness of using HMMs to correct for measurement error in structural parameter estimates, since, in particular cases, they allow for the use of linked data with relatively low sensitivity to linkage error. This is especially appealing, as the methods available to correct for linkage error often cannot be easily applied in this context.

A disadvantage of our findings is that, since linkage error may be absorbed into measurement error parameters, these parameters no longer give "pure" estimates of measurement error. In other words, when the measurement, and not the structural, parameters are of primary interest (e.g., Biemer 2011), our results suggest that linkage and measurement error will be partially conflated. Considering the increasing use of HMMs for this goal, future work should therefore develop methods to correct latent variable model estimates for linkage error, perhaps by extending the estimating equations approach discussed in Chambers and Kim (2016).

Furthermore, while our manuscript provided novel results on the effect of linkage error on point estimates, the effect on the variance of these estimates remains unknown. For false-negative linkage errors (i.e., missed links), the standard theory of missing data applies, and the observed information will equal the information without these errors minus the information that would have been obtained in the missed links (Little and Rubin 2002, p. 191). The effect of false-positive links (i.e., incorrectly linked records) on the variance, however, remains an open question for future work.

SUPPLEMENTARY MATERIALS

Supplementary materials are available online at academic.oup.com/jssam.

References

Alwin, D. F. (ed.) (2007), *Margins of Error: A Study of Reliability in Survey Measurement* (Vol. 547), Hoboken, NJ: John Wiley & Sons.

Alwin, D. F., E. M. Baumgartner, and B. A. Beattie (2017), "Number of Response Categories and Reliability in Attitude Measurement," *Journal of Survey Statistics and Methodology*, 6, 212–239.

Ariel, A., B. Bakker, M. de Groot, G. van Grootheest, J. van der Laan, J. Smit, and B. Verkerk (2014), "Record Linkage in Health Data: A Simulation Study," Statistics Netherlands Discussion Paper, Available at https://www.cbs.nl/nl-nl/achtergrond/2014/16/record-linkage-in-health-data-a-simulation-study. Accessed April 5, 2019.

Armstrong, J., and J. Mayda (1993), "Linkage Error Rates," *Survey Methodology*, 19, 137–147.

Bakker, B. F., and P. J. Daas (2012), "Methodological Challenges of Register-Based Research," *Statistica Neerlandica*, 66, 2–7.

Bassi, F., J. A. Hagenaars, M. A. Croon, and J. K. Vermunt (2000), "Estimating True Changes When Categorical Panel Data Are Affected by Uncorrelated and Correlated Classification Errors: An Application to Unemployment Data," *Sociological Methods & Research*, 29, 230–268.

Biemer, P. (2004), "An Analysis of Classification Error for the Revised Current Population Survey Employment Questions," *Survey Methodology*, 30, 127–140.

Biemer, P. (ed.) (2011), *Latent Class Analysis of Survey Error* (Vol. 571), Hoboken, NJ: John Wiley & Sons.

Biemer, P., E. De Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West (eds.) (2017), *Total Survey Error in Practice*, Hoboken, NJ: John Wiley & Sons.

Biemer, P., and S. L. Stokes (2004), "Approaches to the Modeling of Measurement Errors," in *Measurement Errors in Surveys* (Vol. 173), eds. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, Hoboken, NJ: John Wiley & Sons.

Billiet, J. B., and E. Davidov (2008), "Testing the Stability of an Acquiescence Style Factor Behind Two Interrelated Substantive Variables in a Panel Design," *Sociological Methods & Research*, 36, 542–562.

Blakely, T., and C. Salmond (2002), "Probabilistic Record Linkage and a Method to Calculate the Positive Predictive Value," *International Journal of Epidemiology*, 31, 1246–1252.

Bohensky, M. A., D. Jolley, V. Sundararajan, S. Evans, D. V. Pilcher, I. Scott, and C. A. Brand (2010), "Data Linkage: A Powerful Research Tool with Potential Problems," *BMC Health Services Research*, 10, 346.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (eds.) (2006), *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.), Boca Raton, FL: CRC Press.

Chambers, R. (2009), "Regression Analysis of Probability-Linked Data," Official Statistics Research Series, 4, Statistics New Zealand, Available at http://www3.stats.govt.nz/statisphere/Official_Statistics_Research_Series/Regression_Analysis_of_Probability-Linked_Data.pdf.

Chambers, R., and G. Kim (2016), "Secondary Analysis of Linked Data," in *Methodological Developments in Data Linkage*, eds. K. Harron, H. Goldstein, and C. Dibben, West Sussex: John Wiley & Sons, Ltd.

Di Consiglio, L., and T. Tuoto (2018), "When Adjusting for the Bias Due to Linkage Errors: A Sensitivity Analysis," *Statistical Journal of the IAOS*, 34, 589–597.

Edwards, S. L., M. E. Berzofsky, and P. Biemer (2017), "Effect of Missing Data on Classification Error in Panel Surveys," *Journal of Official Statistics*, 33, 551–570.

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

Fienberg, S. E., and J. P. Gilbert (1970), "The Geometry of a Two by Two Contingency Table," *Journal of the American Statistical Association*, 65, 694–701.

Fuller, W. A. (ed.) (1987), *Measurement Error Models*, (Vol. 204), Hoboken, NJ: John Wiley & Sons.

Galimard, J.-E., S. Chevret, C. Protopopescu, and M. Resche-Rigon (2016), "A Multiple Imputation Approach for MNAR Mechanisms Compatible with Heckman's Model," *Statistics in Medicine*, 35, 2907–2920.

Georgiadis, M. P., W. O. Johnson, I. A. Gardner, and R. Singh (2003), "Correlation-Adjusted Estimation of Sensitivity and Specificity of Two Diagnostic Tests," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 63–76.

Goldstein, H., K. Harron, and A. Wade (2012), "The Analysis of Record-Linked Data Using Multiple Imputation with Data Value Priors," *Statistics in Medicine*, 31, 3481–3493.

Hagenaars, J. A. (1988), “Latent Structure Models with Direct Effects between Indicators: Local Dependence Models,” *Sociological Methods & Research*, 16, 379–405.

———. (ed.) (1990), *Categorical Longitudinal Data: Log-Linear Panel, Trend, and Cohort Analysis*, Newbury Park, CA: Sage Publications.

Harron, K. L., J. C. Doidge, H. E. Knight, R. E. Gilbert, H. Goldstein, D. A. Cromwell, and J. H. van der Meulen (2017), “A Guide to Evaluating Linkage Quality for the Analysis of Linked Data,” *International Journal of Epidemiology*, 46, 1699–1710.

Jones, G., W. O. Johnson, T. E. Hanson, and R. Christensen (2010), “Identifiability of Models for Multiple Diagnostic Testing in the Absence of a Gold Standard,” *Biometrics*, 66, 855–863.

Kim, G., and R. Chambers (2012a), “Regression Analysis under Incomplete Linkage,” *Computational Statistics & Data Analysis*, 56, 2756–2770.

———. (2012b), “Regression Analysis Under Probabilistic Multi-Linkage,” *Statistica Neerlandica*, 66, 64–79.

Kuha, J., and C. Skinner (1997), “Categorical Data Analysis and Misclassification,” in *Survey Measurement and Process Quality* (Vol. 324), eds. L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin, Hoboken, NJ: John Wiley & Sons.

Lahiri, P., and M. D. Larsen (2005), “Regression Analysis with Linked Data,” *Journal of the American Statistical Association*, 100, 222–230.

Leroux, B. G. (1992), “Maximum-Likelihood Estimation for Hidden Markov Models,” *Stochastic Processes and Their Applications*, 40, 127–143.

Liseo, B., and A. Tancredi (2011), “Bayesian Estimation of Population Size Via Linkage of Multivariate Normal Data Sets,” *Journal of Official Statistics*, 27, 491–505.

Little, R. J., and D. B. Rubin (2002), *Statistical Analysis with Missing Data* (Vol. 333, 2nd ed.), Hoboken, NJ: John Wiley & Sons.

Marshall, A., D. G. Altman, P. Royston, and R. L. Holder (2010), “Comparison of Techniques for Handling Missing Covariate Data within Prognostic Modelling Studies: A Simulation Study,” *BMC Medical Research Methodology*, 10, 7.

McLachlan, G., and T. Krishnan (2008), *The EM Algorithm and Extensions* (Vol. 382, 2nd ed.), Hoboken, NJ: John Wiley & Sons.

Oberski, D. L. (2016), “Beyond the Number of Classes: Separating Substantive from Non-Substantive Dependence in Latent Class Analysis,” *Advances in Data Analysis and Classification*, 10, 171–182.

Oberski, D. L., J. A. Hagenaars, and W. E. Saris (2015), “The Latent Class Multitrait-Multimethod Model,” *Psychological Methods*, 20, 422–443.

Oberski, D. L., A. Kirchner, S. Eckman, and F. Kreuter (2017), “Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models,” *Journal of the American Statistical Association*, 112, 1477–1489.

Pankowska, P., B. Bakker, D. L. Oberski, and D. Pavlopoulos (2018), “Reconciliation of Inconsistent Data Sources by Correction for Measurement Error: The Feasibility of Parameter Re-Use,” *Statistical Journal of the IAOS*, 34, 317–329.

Pavlopoulos, D., and K. J. Vermunt (2015), “Measuring Temporary Employment. Do Survey or Register Data Tell the Truth?,” *Survey Methodology*, 41, 197–214.

Qu, Y., and A. Hagdu (2012), “Modeling Correlations between Diagnostic Tests in Efficacy Studies,” in *Modelling Longitudinal and Spatially Correlated Data* (Vol. 122), eds. T. G. Gregoire, D. R. Brillinger, P. Diggle, E. Russek-Cohen, W. G. Warren, and R. D. Wolfinger, New York: Springer Science & Business Media.

Sadinle, M. (2014), “Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach,” *The Annals of Applied Statistics*, 8, 2404–2434.

———. (2017), “Bayesian Estimation of Bipartite Matchings for Record Linkage,” *Journal of the American Statistical Association*, 112, 600–612.

Sadinle, M., R. Hall, and S. E. Fienberg (2011), “Approaches to Multiple Record Linkage,” *Proceedings of International Statistical Institute*, 260, 1–20.

Saris, W. E., and I. N. Gallhofer (2007), *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (2nd ed.), Hoboken, NJ: John Wiley & Sons.

Steorts, R. C. (2015), “Entity Resolution with Empirically Motivated Priors,” *Bayesian Analysis*, 10, 849–875.

Steorts, R. C., R. Hall, and S. E. Fienberg (2016), “A Bayesian Approach to Graphical Record Linkage and Deduplication,” *Journal of the American Statistical Association*, 111, 1660–1672.

Torrance-Rynard, V. L., and S. D. Walter (1997), “Effects of Dependent Errors in the Assessment of Diagnostic Test Performance,” *Statistics in Medicine*, 16, 2157–2175.

Vacek, P. M. (1985), “The Effect of Conditional Dependence on the Evaluation of Diagnostic Tests,” *Biometrics*, 41, 959–968.

Vermunt, J. K., and J. Magidson (2002), “Latent Class Cluster Analysis,” in *Applied Latent Class Analysis*, eds. J. Hagenaars and A. McCutcheon, Cambridge: Cambridge University Press.

———. (2013), *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*, Belmont, MA: Statistical Innovations Inc.

Winglee, M., R. Valliant, and F. Scheuren (2005), “A Case Study in Record Linkage,” *Survey Methodology*, 31, 3–11.

Winkler, W. E. (1999), “The State of Record Linkage and Current Research Problems,” Statistical Research Division, U.S. Census Bureau, Available at https://www.census.gov/srd/papers/pdf/rr99-04.pdf. Accessed April 5, 2019.

Appendix

Appendix A.1. The Effect of Local Independence Assumption Violations on HMM Estimates—An Illustration Using Real Data

As mentioned in the introduction, the local independence assumption, which is necessary for model identification for the standard, one-indicator HMM, is in many cases unrealistic for both survey and register data. If this assumption is violated, HMM estimates are likely to suffer from (considerable) bias and, as such, it is necessary to relax it, which is possible when using multiple indicators per time point.

We provide here an illustration of the biasing effects of local independence assumption violations using data on labor mobility in the Netherlands from the Employment Register (ER) and the Labour Force Survey (LFS) for the years 2009 and 2010. In doing so, we compare the temporary to permanent employment transition estimates obtained using a one-indicator HMM that only uses register data to those obtained using a two-indicator HMM applied to linked ER and LFS data. In the latter model specification, we relax the local independence assumption for the register data, as it is known from previous research (Pavlopoulos and Vermunt 2015; Pankowska et al. 2018) that the measurement error in ER is autocorrelated and, thus, that the local independence assumption is violated. Table A.1 compares the transition estimates obtained from both models; as can be seen, the one-indicator HMM, which

Table A.1. Transition Estimates and Bias for One- and Two-Indicator HMMs

Model specification	Transition estimate (temp → perm)	Absolute bias	Relative bias
One-indicator HMM • Only using ER • Retaining ICE	0.0689	0.0521	310%
Two-indicator HMM • Using ER and LFS • Relaxing ICE for ER	0.0168	–	–

For computational reasons, the simulation study uses a two-indicator HMM that does not relax the local independence assumption for the register data and does not model autocorrelated measurement error in the ER; therefore, the transition rate in the absence of linkage error resembles the one obtained from a single indicator HMM in Appendix A.1.

erroneously retains the local independence assumption for the register data, significantly overestimates the transition rate from temporary to permanent employment. The relative bias resulting from ignoring the violation of the local independence assumption amounts to 310 percent.

Appendix A.2. Fitting of a Latent Class Model to Data with Independent Linkage Error—A Geometric Argument

Jones, Johnson, Hanson, and Christensen (2010) adapted the geometric approach of Fienberg and Gilbert (1970) to the analysis of cross-tables, in order to depict maximum likelihood estimation of the measurement parameters and the structural parameter π in a three-indicator LCM. Here we demonstrate how these estimates are affected by independent linkage error. In the Fienberg and Gilbert (1970) approach, all possible normalized 2×2 cross-tables are placed in a tetrahedron representing the simplex $\{x \in R^4 : \sum x_i = 1\}$ (figure A.1). The four corners of this tetrahedron, A_1, A_2, A_3, and A_4, correspond to cross-tables with all probability mass in a single cell; all other 2×2 cross-tables can be represented as a single point within the tetrahedron. An important subset of tables is the "independence surface" formed by all 2×2 independence tables, which is shown in figure A.1 as the shaded surface. Points along a line on this surface correspond to all independence tables with constant row or column margins.

Following Jones et al. (2010), we consider a binary latent class model with three binary indicators Y_1, Y_2, and Y_3. Without loss of generality, we consider the bivariate cross-table of Y_1 and Y_2 given $Y_3 = 0$ (point p_0) and $Y_3 = 1$ (point p_1). The maximum-likelihood estimates of the conditional distributions

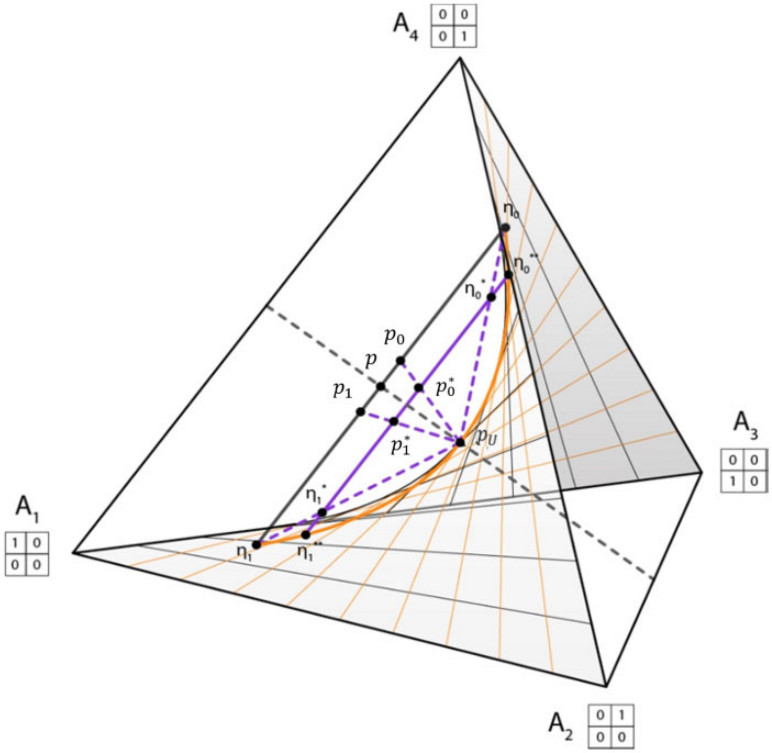

Figure A.1. Geometrical View of Fitting of a Latent Class Model to Data with Independent Linkage Error.

given the latent class $P(Y_1, Y_2|X = 0) = \eta_0$ and $P(Y_1, Y_2|X = 1) = \eta_1$ are then found at the two intersections of the "solution line" $p_1 - p_0$ with the independence surface. This follows from the latent class model's assumption that p_0 and p_1 are both convex combinations of η_0 and η_1, which, by conditional independence given the latent class variable, X, must lie on the independence surface. The MLE of $\mathrm{P}(X|Y_3 = 0)$ is then found as $1 - \mathrm{length}(p_0 - \eta_0)/\mathrm{length}(\eta_1 - \eta_0)$ and, similarly, $\hat{\mathrm{P}}(X|Y_3 = 1) = 1 - \mathrm{length}(p_1 - \eta_1)/\mathrm{length}(\eta_1 - \eta_0)$, implying the MLE for π can be found by applying Bayes's rule (Jones et al. 2010). Note that the length of the line segment $\eta_1 - \eta_0$ indicates the overall accuracy; as η_0 and η_1 lie at greater distance from each other, accuracy estimates under the LCM increase, with the maximum attained at the corners of the tetrahedron (estimated sensitivity and specificity equal to one).

We now consider how the MLEs are affected by independent linkage error. Following (6), we consider the distribution of linked records as a mixture over true matches and true non-matches, indicated by a random variable U. When false-positive linkage error is independent, $P(Y|U) = P(Y)$, the $P(Y_j)$ in the equation above reduce to the marginals under the model, $P(Y_j) = \sum_{y_{k\neq j}} P_{\mathrm{HMM}}(Y)$. This point, p_U in figure A.1, can be found by projecting the marginal over Y_3, point p, onto the independence surface along the line perpendicular to A_1A_2 and A_2A_3 (Fienberg and Gilbert 1970, p. 699). The linkage error model in (6) then shows that the joint distribution under linkage error is a convex combination of p_U and the original joint distribution. That is, under independent linkage error, p_0 and p_1 are "shrunk" toward p_U by exactly $P(U)$. Therefore, when linkage error is independent, the observed data points p_0^* and p_1^* lie on a solution line parallel to the original solution line, with length $(p^* - p)$/length $(p - p_U) = P(U)$.

Similarly, the "true" measurement parameters η_0^* and η_1^* are also convex combinations with p_U, as shown in figure A.1 by points on the line segments $\eta_0 - p_U$ and $\eta_1 - p_U$. Thus, under independence, η_0^* and η_1^* must move closer to p_U and away from the corners of the tetrahedron that represent perfect measurement, shortening the overall length of the solution line. In other words, independent linkage error necessarily leads to higher classification errors. The MLEs of these measurement parameters, η_0^{**} and η_0^{**}, meanwhile, are found by projecting the solution line, not onto $\eta_0 - p_U$ and $\eta_1 - p_U$, but rather onto the independence surface. The distances length$(\eta_0^* - \eta_0^{**})$ and length$(\eta_1^* - \eta_1^{**})$ reflect violations of the LCM's conditional independence assumption. Therefore, linkage error does cause violations of the model's assumptions. However, as can be seen in figure A.1, these violations will be negligible in practice, and the bias is bounded by a small number (relative to the solution line) that depends on $P(U)$. In short, independent linkage errors are absorbed by the measurement parameters, leaving the structural parameters approximately unaffected.

In contrast, bias will be strong when linkage error is not independent, $P(Y|U) \neq P(Y)$. In this case, the new point may lie anywhere on the independence surface, destroying the parallel property of the new solution line. In this case, none of the previous results apply, and the bias in both measurement and structural parameters can be arbitrarily large.

Finally, we have assumed that the mislinked records have an independent joint distribution. When this assumption does not hold, the projection p_U should be replaced by a projection, $p_{U,\mathrm{dep}}$, say, onto a "dependence surface" defined by a constant odds ratio (Fienberg and Gilbert 1970, pp. 699–701). Because of independence of linkage errors, the projection will still be orthogonal to A_1A_2 and A_2A_3. In this situation, the length of the solution line will still be reduced and classification errors will rise. However, the distance from the "true" interpolation between p_U^* and η to the corresponding projection onto the independence surface may increase. In other words, in this situation, depending on the strength of the dependence p_U^*, some non-negligible bias in the MLE of π may start to occur. In particular, for positive dependence (odds ratio > 1), π will be somewhat underestimated (overestimated for negative dependence).

In this appendix, we have indicated the consequences of linkage error for latent class analysis, and have argued that independent linkage errors lead to a relatively small violation of the LCM's assumptions. Although we have not shown this here, we conjecture that the argument extends to higher-dimensional and multiple-category problems, such as the HMM. We have also seen that dependence of linkage errors has more potential to cause bias than dependence in the mislinked records. Our paper investigates these conjectures using a simulation study.

Appendix A.3. The Combined LFS and ER Dataset

A.3.1 Background Information on the LFS and ER

The Labour Force Survey (LFS) is an address-based sampling survey conducted by Statistics Netherlands, which provides information on individuals' labor market position. As of the last quarter of 1999, it has been a rotating panel survey that consists of five waves conducted every three months.

The Employment Register (ER) is an administrative dataset managed by the Dutch Employee Insurance Agency (UWV). It contains monthly information on wages, benefits, and labor relations and covers all insured employees in the Netherlands. While the dataset combines information from various sources, the core information is delivered by employers to the Dutch Tax Authorities (in Dutch: Belastingdienst) for tax purposes. The data from both the LFS and the ER are linked at the individual level to the Population Register (PR), and so the target population of the data is restricted to individuals registered in the Netherlands.

A.3.2 Missing Values

The dataset is unbalanced for the LFS, as it suffers from attrition and has, for the non-survey months, observations missing completely at random (MCAR). More specifically, the first wave of the survey includes 8,708 individuals (130,620 observations), the second 7,458 (111,870 observations), the third 6,856 (102,840 observations), the fourth 6,739 (101,085 observations), and the fifth 6,560 (98,400 observations). While ostensibly the ER cannot suffer from dropout, as all employers are obliged by law to submit their reports, 2,619 observations are missing, which amounts to just under 2 percent of the sample. Those observations are also assumed to be MCAR.

A.3.3 Record Linkage Procedure

The data from both sources are linked at the individual level to the PR. For the LFS, the linkage key is the combination of birth date, gender, postal code, and house number. In the first step, two records are linked if the post code and house number correspond and only one of the other variables of the linkage key differs. In the second step, the remaining, unlinked records are linked on postal code, birth date, and gender, and no differences on the other variables are allowed. This results in a linkage effectiveness, that is, the percentage of linked records, of 98.3 percent for those who had a first interview in 2009.

The ER is linked to the PR in three steps; the procedure is repeated monthly, and one-to-one matching is enforced. In the first step, the records from both sources are linked on the Citizen Service Number (BSN; a unique personal number allocated to everyone registered in the Netherlands). For those records that are linked in this step, it is verified whether birth date and gender are consistent in both data sources. If not, the records go to the next step together with those that were not linked on BSN. In the second step, the data are linked using birth date, gender, postal code, and house number. In the third step, the remaining unlinked records from the first two steps are linked using only the BSN, ignoring any differences in the other variables. This procedure is repeated monthly. The overall linkage effectiveness is approximately 96–97 percent, depending on the chosen month; 99.8 percent of all linked records are successfully linked in the first step.

The linkage to the Population Register results in the assignment of a meaningless linkage number to each linked record of both sources. That linkage number can be used to combine the LFS and ER as well as the data from the successive follow-ups. Having selected only individuals aged 25–55, the linkage effectiveness of the combined sources is approximately 97 percent. The unlinked records refer to cross-border workers from Belgium or Germany that belong to the target population of the ER but not the LFS, as well as to nonregistered individuals (typically immigrants) that are represented in the LFS but not in the ER. Therefore, when focusing on the population of registered individuals that reside in the Netherlands, the linkage of the two data sources of our dataset approaches perfection.

Appendix A.4. Simulation Design

The simulations are designed in the following way. First, we identify young individuals or individuals who had at least one three-monthly transition from temporary to permanent employment recorded in the register data (this step is skipped for random mislinkage conditions). Second, we assign one of two exclusion/mislinkage probabilities to each individual: a higher one for individuals identified in the first step (i.e., younger or who have "transitioned") and a lower one for all remaining ones (we assign the same probability to everyone in the random mislinkage conditions). Third, given the assigned probabilities, we select individuals for exclusion/mislinkage at random. Fourth, in the case of false-negative linkage conditions, we exclude the chosen individuals; in false-positive linkage conditions, we assign the selected individuals to a donor and replace their ER contract type with that of the donor. The assignment to the donor can be either completely random or based on similarity given the age, gender, nationality, and education of individuals. Finally, we run our HMM on the simulated datasets and compare the estimated transitions rates to those obtained when no linkage error is introduced into the dataset.

Below we provide pseudocodes illustrating the simulation design. Both pseudocodes illustrate conditions characterized by an overall 5 percent error rate and in which individuals who have transitioned (from temporary to permanent employment according to the register data) are oversampled.

A.4.1 Pseudocode for a False-Negative Linkage Error Condition

```
Step 1
    1. Identify individuals who have had one or more three-monthly transitions:
           $Temp_{t-3} \rightarrow Perm_t$
2. If a given individual has had a transition, set their exclusion threshold $t$ to 0.15
    a. Else, assign threshold $t$ to 0.05
Step 2
    3. For each individual in the sample, draw a random number from a
       standard uniform distribution—$U_i \sim U(0, 1)$
    4. If $U_i \leq t$, exclude individual $i$
        a. Else, do not exclude individual $i$
Step 3
    5. Run the HMM on this new dataset and compare the results to the original
       ones
```

A.4.2 Pseudocode for a False-Positive Linkage Error Condition

```
Step 1
    1. Identify individuals who have had one or more three-monthly
       transitions:
           $Temp_{t-3} \rightarrow Perm_t$
    2. If a given individual has had a transition, assign mislinkage threshold $t$ as
       0.15
        a. Else, assign threshold $t$ as 0.05
Step 2
    3. For each individual in the sample, draw a random number from a standard
       uniform distribution—$U_i \sim U(0, 1)$
    4. If $U_i \leq t$, mislink individual $i$
        a. Else, do not mislink individual $i$
If the donor is random:
```

5. Assign to the linkage recipient the ER contract type of a randomly chosen individual

If the donor is based on characteristics:

5. a. Use R's matchit package to perform statistical matching based on age, gender, nationality, and education
   b. Assign to the linkage recipient the ER contract type of the matched individual

Step 3

6. Run the HMM on this mislinked data and compare the results to the original ones

Appendix A.5. Illustration of Simulation Results

Figures A.2 and A.3 provide an illustration of the relationship between the type (random, age, or transition-dependent), level (5 percent, 10 percent, 20 percent), and bias introduced by false-negative and false-positive linkage error, respectively.

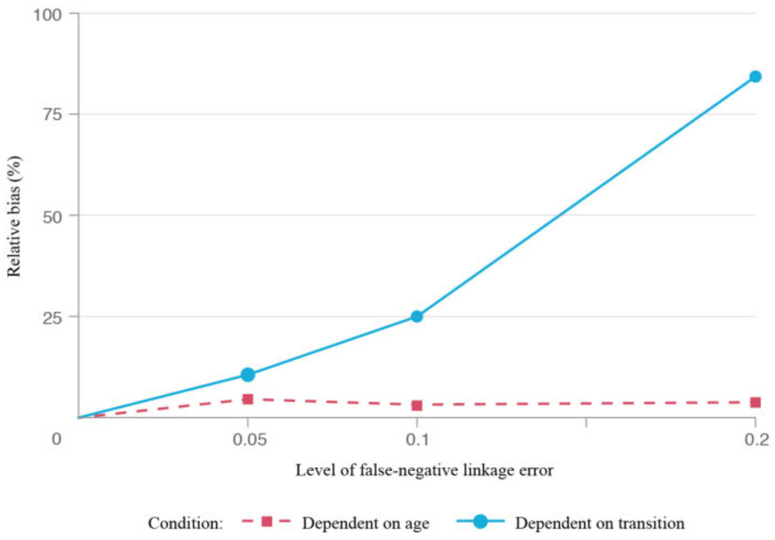

Figure A.2. Relative Bias by Overall Level of False-Negative Linkage Error.

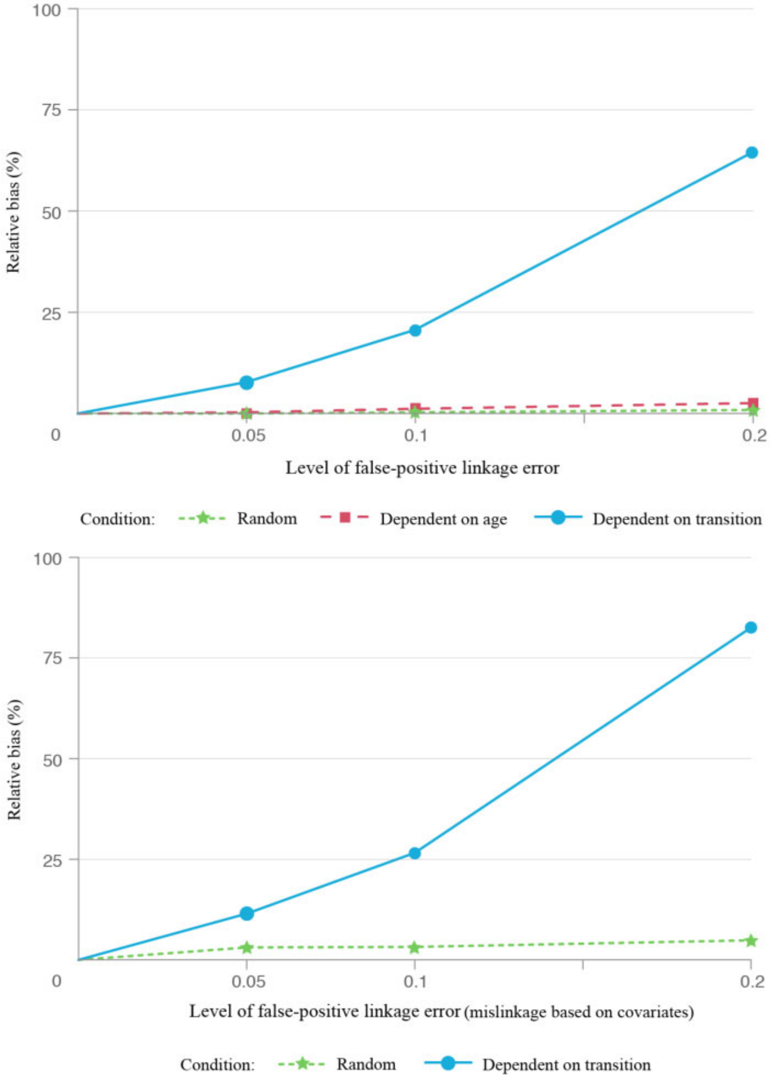

Figure A.3. Relative Bias by Overall Level of False-Positive Linkage Error.