Received: 8 April 2021 | Accepted: 13 June 2022

DOI: 10.1111/rssa.12902

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

Measuring risk of re-identification in microdata: State-of-the art and new directions

Natalie Shlomo[1] | **Chris Skinner**[2,*,†]

[1]Social Statistics Department, School of Social Sciences, University of Manchester, Manchester, UK

[2]Department of Statistics, London School of Economics and Political Sciences, London, UK

Correspondence
Natalie Shlomo, Social Statistics Department, School of Social Sciences, University of Manchester, Humanities Bridgeford Street, Manchester, UK.
Email: Natalie.shlomo@manchester.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Isaac Newton Institute for Mathematical Sciences, Grant/Award Number: EP/K032208/1

Abstract

We review the influential research carried out by Chris Skinner in the area of statistical disclosure control, and in particular quantifying the risk of re-identification in sample microdata from a random survey drawn from a finite population. We use the sample microdata to infer population parameters when the population is unknown, and estimate the risk of re-identification based on the notion of population uniqueness using probabilistic modelling. We also introduce a new approach to measure the risk of re-identification for a subpopulation in a register that is not representative of the general population, for example a register of cancer patients. In addition, we can use the additional information from the register to measure the risk of re-identification for the sample microdata. This new approach was developed by the two authors and is published here for the first time. We demonstrate this approach in an application study based on UK census data where we can compare the estimated risk measures to the known truth.

KEYWORDS

disclosure risks, key variables, log-linear models, model specification, probability scores estimation, registers

[*]Passed away on 21 February 2020.
[†]Authorship for Section 3.

1 | INTRODUCTION

Among Chris Skinner's many influential areas of interest that had a large impact on the dissemination of official statistics was statistical disclosure control (SDC), particularly focusing on measuring the risk of re-identification in sample microdata where the sample is drawn randomly from a finite population. This research was largely motivated by his collaborations with researchers at the University of Manchester in the early 90s to convince the Office of Population Censuses and Surveys (which later merged with the Central Statistical Office to form the Office for National Statistics in 1996) to release a small sample of anonymized records from the 1991 census. This led to the important paper Skinner et al. (1994) which resulted in the release of the Sample of Anonymised Records (SARs) in the United Kingdom for every census since 1991. It was during this time that Chris and others researched estimating the risk of re-identification of sample microdata by developing theory and statistical modelling frameworks, and conceptualized the disclosure risk in terms of population uniqueness given the observed sample microdata (Duncan & Lambert, 1989; Paass, 1988; Skinner, 1992).

The disclosure risk scenario for the release of sample microdata containing records from a survey where the sample is drawn randomly from a finite population is based on the following assumptions: (1) there is an 'intruder' (someone with malicious intent to discredit the statistical office) who has access to the microdata and other auxiliary information from the population that allows him/her to link data sources in order to identify individuals in the sample microdata; (2) there is no 'response knowledge' meaning that the intruder does not know who was drawn into the sample of the survey. The basic definition of the risk of re-identification is therefore the probability of correctly being able to make this match. If the characteristics of the population are known, such as measured in a population register or census, this probability would be relatively straightforward to calculate. However, this is rarely the case since within statistical agencies, samples are typically drawn from area or address-based sample frames. Therefore, a statistical modelling framework is needed to estimate the probability of re-identification. This probability is conditional on the released data and information available to the intruder and defined with respect to a probabilistic model and assumptions about how the data are generated (knowledge of the sampling process). The model is with respect to key variables defined as a set of quasi-identifiers in both data sources, typically categorical such as age, sex, location, ethnicity, that when cross-classified can be used to identify cells with small sample sizes, and we particularly focus on the sample uniques. The risk of re-identification is based on the notion of population uniqueness on the set of key variables: given an observed sample unique in a table generated from the key variables, what is the probability that the cell is also a population unique?

The probabilistic modelling to estimate population uniqueness from the observed sample microdata was developed under two approaches: a full model-based framework taking into account all of the information available to intruders and modelling their behaviour (Duncan & Lambert, 1989, Lambert, 1993 and later Reiter, 2005) and a more simplified approach that restricts the information that would be known to intruders (Benedetti et al., 1998; Bethlehem et al., 1990; Fienberg & Makov, 1998; Skinner & Holmes, 1998).

In Section 2, we provide an overview of the research that Chris Skinner carried out on the probabilistic modelling approach to estimate the risk of re-identification based on population uniqueness. In Section 3 we present new research that was progressed during the Data Linkage and Anonymization Programme at the Isaac Newton Institute, Cambridge UK from July to December 2016, some of which is published here for the first time based on the joint collaboration between the authors. The research extends the previous work of measuring the

risk of re-identification in sample microdata to measuring the risk of re-identification in a publically available register containing a subpopulation where the membership is unknown and may be sensitive, such as a register of cancer patients. The register is clearly not a random sample of the population and hence cannot be used in the original framework described in Section 2. Section 4 presents results from an application study. In Section 5 we present conclusions and future research directions for extending the probabilistic modelling framework to measure the risk of re-identification in non-probability samples which are becoming more prevalent in recent years.

2 | MEASURING THE RISK OF RE-IDENTIFICATION IN SAMPLE MICRODATA

In any released sample microdata from surveys of households and individuals, the direct identifying key variables, such as name, address or identification numbers, are removed. Nevertheless, disclosure risks can arise when there are small counts on a set of cross-classified indirect identifying key variables which are typically categorical, such as: age, sex, place of residence, marital status and occupation, as these can be used to identify an individual and further confidential information may be learnt from survey target variables. Under a probabilistic modelling approach, disclosure risk is assessed on the contingency table of sample counts spanned by these identifying key variables. The assumption is that the sample microdata contain responding individuals in a survey and the population counts are unknown (or only partially known through some marginal distributions). The risk of re-identification is therefore a function of both the population and the sample, and in particular the cell counts of the contingency table. Shlomo (2010) provides an overview of disclosure risk assessment in sample microdata which is summarized here to emphasize the contributions by Chris Skinner.

Individual per-record risk measures in the form of a probability of re-identification are estimated. These per-record risk measures are then aggregated to obtain global risk measures for the entire file. We denote F_k the population size in cell k of a table spanned by key variables having K cells, f_k the sample size in cell k, $\sum_k F_k = N$ and $\sum_k f_k = n$. The set of sample uniques is defined by: $SU = \{k : f_k = 1\}$ and these are the high-risk records with the potential to be population uniques. Two global disclosure risk measures (where I is the indicator function) are the following:

1. Number of sample uniques that are population uniques:

$$\tau_1 = \sum_k I\left(f_k = 1,\ F_k = 1\right)$$

2. Expected number of correct matches for sample uniques assuming a random assignment within cell k. For example, if a sample unique matches to three individuals in the population, the match probability for that sample unique would be 1/3. Aggregating all match probabilities over the sample uniques leads us to: $\tau_2 = \sum_k I\left(f_k = 1\right) 1/F_k$.

If the population frequencies F_k are known then we can easily calculate the global disclosure risk measures. However, this is rarely the case and we assume that the population frequencies F_k are unknown and need to be estimated. Using a probabilistic model, the risk measures are estimated by:

$$\hat{\tau}_1 = \sum_k I(f_k = 1)\,\hat{P}(F_k = 1|f_k = 1) \text{ and } \hat{\tau}_2 = \sum_k I(f_k = 1)\,\hat{E}(1/F_k|f_k = 1). \quad (1)$$

Skinner and Holmes (1998) and Elamir and Skinner (2006) propose a Poisson distribution and a log-linear model to estimate disclosure risk measures in (1). In this model, they assume that $F_k \sim Pois(\lambda_k)$ for each cell k. A sample is drawn by Poisson or Bernoulli sampling with a sampling fraction π_k in cell k: $f_k|F_k \sim Bin(F_k, \pi_k)$. It follows that:

$$f_k \sim \text{Pois}(\pi_k\lambda_k) \text{ and } F_k|f_k \sim \text{Pois}(\lambda_k(1-\pi_k)), \quad (2)$$

where the population cell counts F_k are assumed independent given the sample cell counts f_k.

The parameters λ_k are estimated using log-linear modelling. The sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k\lambda_k$. A log-linear model for the μ_k is expressed as: $\log(\mu_k) = \boldsymbol{x}_k'\boldsymbol{\beta}$ where $\boldsymbol{x_k}$ is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood (MLE) estimator $\hat{\beta}$ are obtained by solving the score equations:

$$\sum_k \left(f_k - \pi_k \exp\left(\boldsymbol{x}'_k\boldsymbol{\beta}\right)\right)\boldsymbol{x_k} = 0 \quad (3)$$

The fitted values are then calculated by: $\hat{\mu}_k = \exp\left(x_k'\hat{\beta}\right)$ and $\hat{\lambda}_k = \hat{\mu}_k/\pi_k$. Individual disclosure risk measures for cell k are:

$$\begin{aligned} P(f_k = 1|f_k = 1) &= \exp(\lambda_k(1-\pi_k)) \\ E(1/F_k|f_k = 1) &= (1 - \exp(\lambda_k(1-\pi_k)))/(\lambda_k(1-\pi_k)) \end{aligned} \quad (4)$$

Plugging $\hat{\lambda}_k$ for λ_k in (4) leads to the estimates $\hat{P}(F_k = 1|f_k = 1)$ and $\hat{E}(1/F_k|f_k = 1)$ and then to $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1). Rinott and Shlomo (2007b) consider confidence intervals for these global risk measures.

Skinner and Shlomo (2008) develop a method for selecting the main effects and interactions for the log-linear model based on estimating and (approximately) minimizing the bias of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. Defining $h(\lambda_k) = P(f_k = 1|f_k = 1)$ for τ_1 and $h(\lambda_k) = E(1/F_k|f_k = 1)$ for τ_2, they consider the expression:

$$B = \sum_k E(I(f_k = 1))\left(h\left(\hat{\lambda}_k\right) - h(\lambda_k)\right).$$

A Taylor expansion of h leads to the approximation

$$B \approx \sum_k \pi_k\lambda_k \exp(-\lambda_k)\left(h'(\lambda_k)\left(\hat{\lambda}_k - \lambda_k\right) + h''(\lambda_k)\left(\hat{\lambda}_k - \lambda_k\right)^2/2\right)$$

and the relations $E(f_k) = \pi_k\lambda_k$ and $E\left(\left(f_k - \pi_k\hat{\lambda}_k\right)^2 - f_k\right) = \pi_k^2 E\left(\hat{\lambda}_k - \lambda_k\right)^2$ under the hypothesis of a Poisson distribution fit lead to a further approximation of B of the form:

$$\hat{B} \approx \sum_k \hat{\lambda}_k \exp\left(-\pi_k\hat{\lambda}_k\right)\left(-h'(\lambda_k)\left(f_k - \pi_k\hat{\lambda}_k\right) + h''(\lambda_k)\left(\left(f_k - \pi_k\hat{\lambda}_k\right)^2 - f_k\right)/(2\pi_k)\right). \quad (5)$$

For example, for τ_1:

$$\hat{B}_1 \approx \sum_k \hat{\lambda}_k \exp\left(-\hat{\lambda}_k\right)(1-\pi_k)\left\{\left(f_k - \pi_k\hat{\lambda}_k\right) + (1-\pi_k)\left[\left(f_k - \pi_k\hat{\lambda}_k\right)^2 - f_k\right] \Big/ (2\pi_k)\right\}. \quad (6)$$

The method selects the model using a forward search algorithm which minimizes the standardized bias estimate $\hat{B}_i/\sqrt{\hat{v}_i}$ for $\hat{\tau}_i, i = 1, 2$, which is used as the goodness-of-fit criteria where $\hat{v}_i$ is the variance estimate of $\hat{B}_i$. The goodness-of-fit criteria $\hat{B}_i/\sqrt{\hat{v}_i}$ have an approximate standard normal distribution under the hypothesis that the expected value of $\hat{B}_i$ is zero.

Skinner and Shlomo (2008) also address the estimation of disclosure risk measures under complex survey designs with stratification, clustering and survey weights. While the method described assumes that all individuals within cell k are selected independently using Bernoulli sampling, that is, $(f_k = 1|F_k) = F_k\pi_k(1-\pi_k)^{F_k-1}$, this may not be the case when sampling clusters (households). In practice, key variables typically include variables such as age, sex and occupation that tend to cut across clusters. Therefore, the above assumption holds in practice in most household surveys and does not cause bias in the estimation of the risk measures. Inclusion probabilities may vary across strata, the most common stratification is on geography. Strata indicators should always be included in the key variables to take into account differential inclusion probabilities in the log-linear model. Under complex sampling, the λ_k can be estimated consistently using pseudo-maximum likelihood estimation (Rao & Thomas, 2003), where the estimating equation in (3) is modified as:

$$\sum_k \left(\hat{F}_k - \exp\left(\boldsymbol{x}'_{\boldsymbol{k}}\boldsymbol{\beta}\right)\right)\boldsymbol{x}_{\boldsymbol{k}} = 0 \quad (7)$$

and $\hat{F}_k$ is obtained by summing the survey weights in cell k: $\hat{F}_k = \sum_{i\in k} w_i$. The resulting estimates $\hat{\lambda}_k$ are plugged into expressions in (4) and π_k is replaced by the estimate $\hat{\pi}_k = f_k/\hat{F}_k$. The goodness-of-fit criteria $\hat{B}$ is also adapted to the pseudo-maximum likelihood approach.

The probabilistic modelling presented here and in other related work in the literature assume that there is no measurement error in the way the data are recorded. Besides typical errors in data capture, key variables can also purposely be misclassified as a means of masking the data, for example through record swapping or the post-randomization method (PRAM) (Gouweleeuw, et al., 1998). Shlomo and Skinner (2010) adapt the estimation of the risk of re-identification of τ_2 in (1) to take into account measurement errors. We denote the cross-classified key variables in the population and the microdata as X and assume that X in the microdata have undergone some misclassification or perturbation error denoted by the value $\tilde{X}$ and determined independently by a misclassification matrix M:

$$M_{kj} = P(\tilde{X} = k|X = j). \quad (8)$$

The record-level disclosure risk measure of a match with a sample unique under measurement error is:

$$\frac{M_{kk}\left(1-\pi_k M_{kk}\right)}{\sum_j F_j M_{kj}/\left(1-\pi_k M_{kj}\right)} \leq \frac{1}{F_k}. \quad (9)$$

Under assumptions of small sampling fractions and small misclassification errors, the disclosure risk measure of τ_2 can be approximated by: $M_{kk}/\sum_j F_j M_{kj}$ or $M_{kk}/\tilde{F}_k$ where $\tilde{F}_k$ is the

population count with $\tilde{X} = k$. Aggregating the per-record disclosure risk measures, the global risk measure is:

$$\tau_2 = \sum_k I\left(f_k = 1\right) M_{kk}/\tilde{F}_k. \tag{10}$$

Note that to calculate the measure in (10) only the diagonal of the misclassification matrix needs to be known, that is, the probabilities of not being perturbed. Population counts are generally not known so the estimate in (10) can be obtained by probabilistic modelling on the misclassified sample as shown above:

$$\hat{\tau}_2 = \sum_k I\left(\tilde{f}_k = 1\right) M_{kk}\hat{E}(1/\tilde{F}_k|\tilde{f}_k). \tag{11}$$

There have been many other contributions expanding the Poisson log-linear modelling framework for estimating the risk of re-identification in survey microdata. Ichim (2008) considers extensions by introducing the survey weights in the analysis of the contingency tables and also proposes a maximum penalized-likelihood approach to obtain smoother estimates of the risk of re-identification. Forster and Webb (2007) extend the log-linear modelling framework to a model averaging approach rather than requiring to choose a single model a priori. They use a Bayesian model averaging technique according to M possible log-linear models but limit the models to decomposable geographical models. The posterior distribution under model uncertainty is hence obtained as a weighted average of the posterior distribution under the various models. Rinott and Shlomo (2006, 2007a) generalize the probabilistic modelling using the Negative Binomial distribution rather than the Poisson distribution and implement the probabilistic modelling framework on local 'neighbourhoods' of the sample uniques. Manrique-Vallier and Reiter (2012) propose an alternative to log-linear models for datasets with sparse contingency tables according to the key variables using a Bayesian version of Grade of Membership (GoM) models and they use a Markov Chain Monte Carlo algorithm for fitting the model. Carota et al. (2015) applied a Bayesian semi-parametric version of log-linear models, specifically a mixed effects log-linear model with a Dirichlet process (DP) prior.

3 | MEASURING THE RISK OF RE-IDENTIFICATION IN A SUBPOPULATION

Up till now, the survey microdata under investigation is a random sample subset of the population and therefore it can be used to estimate population parameters for the probabilistic models to estimate disclosure risk measures based on the risk of re-identification. In addition, the intruder knows that it is possible that an individual in the sample microdata can be matched to the population as all have a non-zero chance of being selected into the sample. In this new setting, we assume that we have microdata that represents a subpopulation. It is publically available but the membership of the subpopulation is not known. For example, a subpopulation can refer to all persons with a medical condition, such as a cancer/HIV register or owns a supermarket loyalty card. The subpopulation is not representative of the population as is the case for a random sample. Similar to the case for sample microdata, we assume the same disclosure risk scenario that an intruder aims to match a record in the subpopulation to an individual in the population of which

the subpopulation is a subset and that the population counts are unknown. We also assume that there are no measurement errors in the way the data are recorded in the register. In order to allow inference about population uniqueness in the subpopulation, we assume that there also exists survey microdata from a random sample and that there are categorical identifying variables X across all data sources that can be used to match a record in the subpopulation/sample microdata to the population.

As mentioned we assume that membership in the subpopulation register, denoted by the variable R, is unknown and may be sensitive and the primary concern is that an intruder can identify an individual in the subpopulation and disclose their value of R. In this case, it is reasonable to assume that the intruder cannot use R as a potential identifying key variable for the probabilistic modelling. Therefore, in order to make inference about population uniqueness the intruder makes use of the sample microdata file where the sample is drawn from the finite population. Note that the membership of the subpopulation R is also not known in the sample microdata. Thus, whereas previously the sample microdata file served two purposes, one as the file about which disclosure risk is a concern and one for inference about population uniqueness, we now suppose that the intruder must resort to using separate files for these two purposes.

As an illustration to this new setting, we assume that both survey microdata containing a random sample from the population, such as the Labour Force Survey microdata, and a Register of Cancer Patients are observed, but the inclusion into the sample and the membership of the register are not known. In addition, it is not known who in the sample is also included in the register. We can then estimate the risk of re-identification in the Register of Cancer Patients where we draw inference from the sample microdata to estimate the population parameters. Alternatively, we can estimate the risk of re-identification in the sample microdata using the additional information that is available in the observed Register of Cancer Patients.

In summary, the key additional complication is that another data source is observed and introduced into the framework as described in Section 2 of a subpopulation register that is not a random subset of the population. This new framework allows for the estimation of the risk of re-identification through population uniqueness for two settings:

- Estimate the risk of re-identification in the subpopulation microdata given the data in both the subpopulation microdata and the sample microdata;
- Estimate the risk of re-identification in the sample microdata given the data in the sample microdata and the additional information that can be obtained from the publically available subpopulation microdata.

3.1 | Framework

Let U and U_1 denote the population and the subpopulation, respectively, with $U_1 \subset U$. We refer to members of U as individuals, although they could more generally be other types of units. Let R_i be the subpopulation indicator variable for individual i with $R_i = 1$ if $i \in U_1$ and $R_i = 0$ otherwise. We suppose that a subpopulation microdata file has been constructed for members of U_1. We are concerned about the possibility of an intruder matching a record in this file to a known individual in the population and thus disclosing the fact that $R_i = 1$ for individual i. As discussed, we suppose that membership R_i is a sensitive variable for which disclosure is undesirable.

We suppose that any matching by the intruder makes use of a vector X of key variables which are included in the subpopulation microdata file and which the intruder may be able to determine

for known individuals in the population. We suppose that R_i is not included in X. We suppose that the key variables are categorical and focus our concern on an intruder who finds an exact match on X between a record in the subpopulation microdata and a known individual in the population (assuming no measurement errors). As before, we label the possible combinations of key variables by $k, k = 1, \ldots, K$ and refer to each combination as a cell in a multi-way contingency table. We denote the population frequency in cell k by F_k so that the cell is population unique if $F_k = 1$.

We denote the subpopulation frequencies in cell k by F_k^1. The most high-risk records are for cells with $F_k^1 = 1$ and, analogous to the derivation presented in Skinner and Shlomo (2008), two alternative risk measures are given by

$$\tau_1^* = \sum_k P(F_k = 1|F_k^1 = 1)I\left(F_k^1 = 1\right) \quad \tau_2^* = \sum_k E(1/F_k|F_k^1 = 1)I\left(F_k^1 = 1\right) \tag{12}$$

We can also convert (12) into proportions by dividing by $\sum_k I\left(F_k^1 = 1\right)$.

There is no way that these measures can be estimated consistently from the subpopulation microdata alone. The microdata provide information about the F_k^1 but not about the F_k in U. We have in mind subpopulations U_1 where the distribution of X may be quite different to that in U so the subpopulation microdata carries no direct information about the F_k. We suppose, therefore, that in addition to the microdata for people in U_1, there is a random sample microdata file in which the values of X are recorded for a probability sample s from U. We suppose that the two microdata files are not linked. Let f_k denote the frequency in cell k in s. Note that the f_k and F_k^1 are observed, but the F_k are not. If the intruder has access to the sample microdata file, then it may be advantageous to restrict attention to cells with $f_k = 1$, leading to the following risk measures

$$\begin{aligned}\tau_1 &= \sum_k P\left(F_k = 1|F_k^1 = 1, f_k = 1\right) I\left(F_k^1 = 1, f_k = 1\right)\\ \tau_2 &= \sum_k E\left(1/F_k|F_k^1 = 1, f_k = 1\right) I\left(F_k^1 = 1, f_k = 1\right)\end{aligned} \tag{13}$$

An alternative approach, following Skinner and Elliot (2002), is to focus on the cells with one entry in cell k of both the subpopulation and sample microdata, where $I\left(F_k^1 = 1, f_k = 1\right)$ represents a sample unique in both sources of microdata, and to note that there are $\sum_k F_k I(F_k^1 = 1, f_k = 1)$ individuals in the population U who could be matched to these individuals using X. An alternative measure of risk is thus given by

$$\theta = \frac{\sum_k I(F_k^1 = 1, f_k = 1)}{\sum_k F_k I(F_k^1 = 1, f_k = 1)} \tag{14}$$

which may be interpreted as the probability that a match is correct if an intruder selects any one of the $\sum_k F_k I(F_k^1 = 1, f_k = 1)$ individuals at random with equal probability.

Following Skinner and Shlomo (2008), suppose that F_k is Poisson distributed, $F_k \sim Pois\,(\lambda_k)$ where the parameter λ_k obeys the log-linear model

$$\log\,(\lambda_k) = \boldsymbol{x}_k'\boldsymbol{\beta}. \tag{15}$$

Suppose that within cell k the unknown membership variable R_i takes the value 1 with probability p_k, independently for each of the F_k units, so that $F_k^1 \sim Pois(\phi_k)$ where $\phi_k = \lambda_k p_k$, and the F_k^1 is binomially distributed $F_k^1 | F_k \sim Bin(F_k, p_k)$ conditional on the F_k. Furthermore, we assume that p_k obeys the logistic model:

$$\text{logit}(p_k) = \boldsymbol{x}_k' \boldsymbol{\xi}. \tag{16}$$

For simplicity, the same vector $\boldsymbol{x}_k$ is used in both models here, but different specifications could apply. Suppose that the sample s is obtained by Poisson or Bernoulli sampling with inclusion probability π_k in cell k.

3.2 | Expressions for risk of re-identification measures

In this section, we provide expressions for the risk measures from Section 2 in terms of the model parameters introduced in Section 3.1. We first introduce more notation. Let f_k^1 denote the frequency in cell k in $s \cap U_1$ and let $\tilde{f}_k^1 = F_k^1 - f_k^1$. Note that $s \cap U_1$ specifies the set of individuals appearing in both the sample and the subpopulation and that this set is not observed. Similarly, let $f_k^2 = f_k - f_k^1$ and let $\tilde{f}_k^2 = (F_k - F_k^1) - f_k^2$. From here we have $\tilde{f}_k^1 \geq 0$, $\tilde{f}_k^2 \geq 0$ and $f_k^1 + f_k^2 = f_k$, $f_k^1 + \tilde{f}_k^1 = F_k^1$ and $f_k^1 + \tilde{f}_k^1 + f_k^2 + \tilde{f}_k^2 = F_k$. Figure 1 shows the Venn diagram of the decomposition of the population count F_k into mutually exclusive sets.

Assuming that the selection of s is independent of R, a convenient approximation for developing the risk measures is to assume that the quantities f_k^1, $\tilde{f}_k^1$, f_k^2 and $\tilde{f}_k^2$ are independent with

$$\begin{aligned} &f_k^1 \sim Pois(\pi_k \phi_k),\ \tilde{f}_k^1 \sim Pois((1-\pi_k)\phi_k),\ f_k^2 \sim Pois(\pi_k(\lambda_k - \phi_k)) - \text{ and} \\ &\tilde{f}_k^2 \sim Pois((1-\pi_k)(\lambda_k - \phi_k)). \end{aligned} \tag{17}$$

To obtain an expression for τ_1 in (13), we write

$$P\left(F_k = 1 | F_k^1 = 1, f_k = 1\right) = \frac{P\left(F_k = 1, F_k^1 = 1, f_k = 1\right)}{P\left(F_k^1 = 1, f_k = 1\right)}. \tag{18}$$

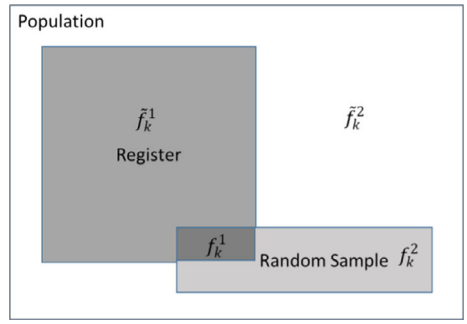

FIGURE 1 Venn diagram of the decomposition of the population counts F_k in each shaded area.

The only possible combination of values $\left(f_k^1, \tilde{f}_k^1, f_k^2, \tilde{f}_k^2\right)$ which leads to $F_k = 1, F_k^1 = 1$ and $f_k = 1$ is given by (1, 0, 0, 0). Hence,

$$\begin{aligned} P\left(F_k = 1 | F_k^1 = 1, f_k = 1\right) &= P\left(f_k^1 = 1\right) P\left(\tilde{f}_k^1 = 0\right) P\left(f_k^2 = 0\right) P\left(\tilde{f}_k^2 = 0\right) \\ &= \pi_k \phi_k \exp\left(-\pi_k \phi_k\right) \times \exp\left(-\left(1-\pi_k\right)\phi_k\right) \times \exp\left(-\pi_k(\lambda_k - \phi_k)\right) \\ &\quad \times \exp\left(-\left(1-\pi_k\right)(\lambda_k - \phi_k)\right). \end{aligned} \tag{19}$$

Furthermore, the only possible combinations of values of $\left(f_k^1, \tilde{f}_k^1, f_k^2,\right)$ which lead to $F_k^1 = 1$ and $f_k = 1$ are given by (1, 0, 0) and (0, 1, 1). Hence

$$\begin{aligned} P\left(F_k^1 = 1, f_k = 1\right) &= P\left(f_k^1 = 1\right) P\left(\tilde{f}_k^1 = 0\right) P\left(f_k^2 = 0\right) \\ &\quad + P\left(f_k^1 = 0\right) P\left(\tilde{f}_k^1 = 1\right) P\left(f_k^2 = 1\right) \\ &= \pi_k \phi_k \exp\left(-\pi_k \phi_k\right) \times \exp\left(-\left(1-\pi_k\right)\phi_k\right) \times \exp\left(-\pi_k(\lambda_k - \phi_k)\right) \\ &\quad + \exp\left(-\pi_k \phi_k\right) \times \left(1-\pi_k\right)\phi_k \exp\left(-\left(1-\pi_k\right)\phi_k\right) \\ &\quad \times \pi_k(\lambda_k - \phi_k) \exp\left(-\pi_k(\lambda_k - \phi_k)\right). \end{aligned} \tag{20}$$

Plugging (19) and (20) into (18) and simplifying gives

$$P\left(F_k = 1 | F_k^1 = 1, f_k = 1\right) = \frac{\pi_k \phi_k \exp\left(-\left(1-\pi_k\right)(\lambda_k - \phi_k)\right)}{\pi_k \phi_k + \left(1-\pi_k\right)\pi_k \phi_k (\lambda_k - \phi_k)} = \frac{\exp\left(-\left(1-\pi_k\right)(\lambda_k - \phi_k)\right)}{1 + \left(1-\pi_k\right)(\lambda_k - \phi_k)}. \tag{21}$$

To evaluate τ_1^*, we use

$$P\left(F_k = 1 | F_k^1 = 1\right) = P\left(F_k - F_k^1 = 0\right) = \exp\left(-(\lambda_k - \phi_k)\right) \tag{22}$$

since $F_k - F_k^1 \sim Pois(\lambda_k - \phi_k)$.

The expression for θ in (14) can be estimated design-consistently without the need for modelling as shown in Section 3.3.

3.3 | Estimation of risk measures

We first consider the estimation of θ in (14). Following the arguments of Skinner and Elliot (2002), a design-consistent estimator of θ is given by

$$\hat{\theta} = \frac{\sum_k \pi_k I(F_k^1 = 1, f_k = 1)}{\sum_k \pi_k I(F_k^1 = 1, f_k = 1) + \sum_k 2\left(1-\pi_k\right) I(F_k^1 = 1, f_k = 2)}, \tag{23}$$

where it is assumed that Poisson or Bernoulli sampling is employed with inclusion probability π_k in cell k.

Design-consistent estimation is not feasible for the remaining risk measures in (12) and (13) and we adopt the probabilistic modelling approach. These measures depend on the unknown λ_k and ϕ_k and the known π_k via (21) and (22). Assuming the models in (15) and (16), the λ_k and ϕ_k are known functions of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. The data consist of the values

f_k and F_k^1. In this section, we consider how to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ from the f_k and F_k^1. We then suppose that the risk of re-identification measures are estimated by plugging these parameter estimates into (21) and (22). If the F_k were observed it would be straightforward to factor the likelihood into two components, one dependent on $\boldsymbol{\beta}$ via $F_k \sim Pois(\lambda_k)$ and (15) and one dependent on $\boldsymbol{\xi}$ via $F_k^1|F_k \sim Bin(F_k, p_k)$ and (16). However, we only observe f_k and not F_k and the conditional distribution $F_k^1|f_k$ cannot be expressed in general as a function of p_k.

We consider a two-step estimation procedure under two different approaches.

3.3.1 | Approach A

In the first approach, Approach A, we first estimate $\boldsymbol{\beta}$ from $f_k \sim Pois(\pi_k\lambda_k)$, combined with the log-linear model defined by (15) as in Skinner and Shlomo (2008). In the second step, we estimate $\boldsymbol{\xi}$, fixing λ_k at the value implied by (15) with $\boldsymbol{\beta}$ set at its value estimated at the first step. We then use (16) and the fact that $\phi_k = \lambda_k p_k$ to write

$$\log\phi_k = \log\lambda_k + \boldsymbol{x'_k}\boldsymbol{\xi} - \log\left(1 + \exp\left(\boldsymbol{x'_k}\boldsymbol{\xi}\right)\right) \tag{24}$$

and then estimate $\boldsymbol{\xi}$ from the fact that $F_k^1 \sim Pois(\phi_k)$ using maximum likelihood estimation and treating λ_k as known. The log likelihood (ignoring a constant term) is given by

$$l(\boldsymbol{\xi}) = \sum_{\boldsymbol{k}}(-\phi_k + F_k^1\log(\phi_k)). \tag{25}$$

The score equations are then given by

$$U(\boldsymbol{\xi}) = \sum_{\boldsymbol{k}}\left(\frac{F_k^1 - \phi_k}{\phi_k}\right)\frac{\partial\phi_k}{\partial\boldsymbol{\xi}} = 0. \tag{26}$$

We obtain from (24) that

$$\frac{\partial\phi_k}{\partial\boldsymbol{\xi}} = \phi_k\frac{\partial\log(\phi_k)}{\partial\boldsymbol{\xi}} = \phi_k\left(\boldsymbol{x'_k} - \frac{\exp\left(\boldsymbol{x'_k}\boldsymbol{\xi}\right)}{1+\exp\left(\boldsymbol{x'_k}\boldsymbol{\xi}\right)}\boldsymbol{x'_k}\right) = \phi_k(1-p_k)\boldsymbol{x'_k}.$$

Hence the score equations can be written as

$$U(\boldsymbol{\xi}) = \sum_{\boldsymbol{k}}(F_k^1 - \phi_k)(1-p_k)\boldsymbol{x'_k} = 0.$$

To obtain an estimator of $\boldsymbol{\xi}$, these equations can be solved by the Newton–Raphson method or the method of Fisher scoring, treating each of ϕ_k and p_k as functions of $\boldsymbol{\xi}$ using (16) and (24) and treating the λ_k as given.

The derivative of $U(\boldsymbol{\xi})$ is

$$\begin{aligned} H(\boldsymbol{\xi}) = \frac{\partial U(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}} &= -\sum_{\boldsymbol{k}}\left((F_k^1 - \phi_k)\frac{\partial p_k}{\partial\boldsymbol{\xi}} + \frac{\partial\phi_k}{\partial\boldsymbol{\xi}}(1-p_k)\right)\boldsymbol{x'_k} \\ &= -\sum_{\boldsymbol{k}}\left((F_k^1 - \phi_k)p_k + \phi_k(1-p_k)\right)(1-p_k)\boldsymbol{x_k}\boldsymbol{x'_k}. \end{aligned}$$

Using the method of Fisher scoring, we replace $H(\xi)$ by its expectation. Thus, using $E(F_k^1) = \phi_k$, we have

$$I(\boldsymbol{\xi}) = E(H(\boldsymbol{\xi})) = -\sum_{\boldsymbol{k}} \phi_k (1-p_k)^2 \boldsymbol{x_k x'_k}$$

and letting $\hat{\xi}_{\boldsymbol{r}}$ denote the estimate of $\boldsymbol{\xi}$ at the rth iteration, and setting $\boldsymbol{\xi}_0 = 0$ we have

$$\hat{\xi}_{\boldsymbol{r}} = \hat{\xi}_{\boldsymbol{r}-1} - I\left(\hat{\xi}_{\boldsymbol{r}-1}\right)^{-1} U\left(\hat{\xi}_{\boldsymbol{r}-1}\right).$$

3.3.2 | Approach B

In the second approach, Approach B, we return to the microdata level and estimate for each individual in the subpopulation file the probability of membership: $\tilde{p}_i$ $i = 1, \ldots, N_1$ where N_1 is the number of individuals in the subpopulation. The estimation is carried out by using the random sample microdata file as the reference sample as described below. We assume that the sample microdata have survey weights for each individual j, w_j, $j = 1, \ldots, n$. Since the probability of membership $\tilde{p}_i$ corrects for the lack of representativeness in the subpopulation, we can calculate estimates for the population totals $\hat{F}_k$ by inverse probability weighted (IPW) estimation: $\hat{F}_k = \sum_{i \in k} 1/\hat{\tilde{p}}_i$ where $\hat{\tilde{p}}_i$ is the estimate of $\tilde{p}_i$. Now treating the $\hat{\tilde{p}}_i$ as fixed from the first step, we then estimate λ_k by the pseudo-maximum likelihood estimation shown in (7) as described in Skinner and Shlomo (2008). Defining $\hat{p}_k = 1/\hat{F}_k$ we estimate $\hat{\phi}_k = \hat{\lambda}_k \hat{p}_k$ and calculate the risk measures in (21) and (22).

To estimate the probability of membership $\tilde{p}_i$ for the subpopulation microdata, we implement the method proposed in Chen et al. (2019) summarized below. We denote the subpopulation microdata as file A and the sample microdata as file B. We stack the two files and define $T_i = 1$ if $i \in A$ and $T_i = 0$ if $i \in B$. The probability of membership for the subpopulation microdata A is $\tilde{p}_i \equiv \tilde{p}_i(\boldsymbol{x_i}, \boldsymbol{\xi}) = P(T_i = 1 | \boldsymbol{x_i}, \boldsymbol{\xi})$ where $\boldsymbol{x_i}$ is the design vector denoting the main effects and interactions. The maximum likelihood estimator of $\tilde{p}_i$ is $\hat{\tilde{p}}_i\left(\boldsymbol{x_i}, \hat{\boldsymbol{\xi}}\right)$ where $\hat{\boldsymbol{\xi}}$ maximizes the log-likelihood function

$$\begin{aligned} l(\boldsymbol{\xi}) &= \sum_{i=1}^{N} \left(T_i \log\left(\tilde{p}_i\right) + (1-T_i)\log\left(1-\tilde{p}_i\right)\right) \\ &= \sum_{i \in A} \log\left(\frac{\tilde{p}_i(\boldsymbol{x_i}, \boldsymbol{\xi})}{1-\tilde{p}_i(\boldsymbol{x_i}, \boldsymbol{\xi})}\right) + \sum_{i=1}^{N} \log\left(1-\tilde{p}_i(\boldsymbol{x_i}, \boldsymbol{\xi})\right). \end{aligned} \tag{27}$$

Since we do not observe the whole population, Chen et al. (2019) replace the second term in (27) with the Horvitz–Thompson estimator obtained from the random reference sample having survey weights w_j and with information on $\boldsymbol{x_i}$, to maximize the pseudo log-likelihood function

$$l^*(\boldsymbol{\xi}) = \sum_{i \in A} \log\left(\frac{\tilde{p}_i(\boldsymbol{x_i}, \boldsymbol{\xi})}{1-\tilde{p}_i(\boldsymbol{x_i}, \boldsymbol{\xi})}\right) + \sum_{j \in B} w_j \log\left(1-\tilde{p}_j\left(\boldsymbol{x_j}, \boldsymbol{\xi}\right)\right). \tag{28}$$

Under a logistic regression model where $\tilde{p}_i \equiv \tilde{p}_i(\boldsymbol{x_i}, \boldsymbol{\xi}) = \frac{exp(\boldsymbol{x'_i \xi})}{1+exp(\boldsymbol{x'_i \xi})}$ the pseudo log-likelihood function is

$$l^*(\boldsymbol{\xi}) = \sum_{i \in A} \boldsymbol{x'_i \xi} - \sum_{j \in B} w_j \log\left(1 + \exp\left(\boldsymbol{x'_j \xi}\right)\right).$$

And the score equations:

$$U^*(\boldsymbol{\xi}) = \frac{\partial l^*(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = \sum_{i \in A} \boldsymbol{x_i} - \sum_{j \in B} w_j \tilde{p}_j \left(\boldsymbol{x_j}, \boldsymbol{\xi}\right) \boldsymbol{x_j} = 0. \tag{29}$$

Chen, et al. (2019) propose a Newton–Raphson procedure. Letting $\hat{\boldsymbol{\xi}}_{\boldsymbol{r}}$ denote the estimate of $\boldsymbol{\xi}$ at the rth iteration, we have

$$\hat{\boldsymbol{\xi}}_{\boldsymbol{r}} = \hat{\boldsymbol{\xi}}_{\boldsymbol{r}-1} - H^*\left(\hat{\boldsymbol{\xi}}_{\boldsymbol{r}-1}\right)^{-1} U^*\left(\hat{\boldsymbol{\xi}}_{\boldsymbol{r}-1}\right),$$

where $H^*(\boldsymbol{\xi}) = \frac{\partial U^*(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} = -\sum_{i \in B} w_i \tilde{p}_i \left(\boldsymbol{x_i}, \boldsymbol{\xi}\right) \left(1 - \tilde{p}_i \left(\boldsymbol{x_i}, \boldsymbol{\xi}\right)\right) \boldsymbol{x_i} \boldsymbol{x'_i}$ and setting $\boldsymbol{\xi}_0 = 0$ for the first iteration.

4 | APPLICATION STUDY

From the UK Census 2001, cell proportions from published tables for ages 16 and over were calculated and cross-classified and if necessary, complemented with iterative proportional fitting, to obtain joint probabilities on the following variables: Geography (6 categories), Age group (14 categories), Sex (2 categories), Marital Status (6 categories), Ethnicity (16 categories), Economic Activity (10 categories) and Ill Health (2 categories). We then multiplied the proportions by 1,000,000 individuals and after rounding obtain a synthetic census microdata dataset of $N = 1{,}003{,}401$. The subpopulation data are those having ill health where $N_1 = 179{,}699$. We produce a multiway contingency table of size $K = 161{,}280$ cells defined by all variables except Ill Health. Table 1 compares the distributions in the population and the subpopulation microdata for key variables Age Group, Sex and Economic Activity. Table 1 clearly shows that the subpopulation mainly contains the elderly population as they are more likely to have Ill Health.

4.1 | Simulation steps approach A

Step 1: Draw 100 random samples without replacement from the population using Bernoulli sampling where $\pi = 1/50$ and resulting in a sample size of $n = 20{,}068$ on average.

Step 2: On each sample, we run a log-linear model (3) where the model is the all two-way interaction model on the key variables: Geography, Age group, Sex, Marital Status, Ethnicity and Economic Activity, to estimate $\hat{\mu}_k = \exp\left(\boldsymbol{x'_k}\hat{\boldsymbol{\beta}}\right)$ and $\hat{\lambda}_k = \hat{\mu}_k/\pi$.

Step 3: Estimate the probability p_k according to Approach A in Section 3.3. Here we use the same key variables and the main effects design matrix.

Step 4: Define $\hat{\phi}_k = \hat{\lambda}_k \hat{p}_k$ and calculate the risk measures in (21) and (22) and compare to the true values based on the known population.

4.2 | Simulation steps approach B

Step 1: Draw 100 random samples without replacement from the population using Bernoulli sampling where $\pi = 1/50$ and resulting in a sample size of $n = 20{,}068$ on average.

TABLE 1 Comparison of distributions in key variables age group, sex and economic activity

	Population U		Subpopulation U_1	
Key variables	**Total**	**Percentage**	**Total**	**Percentage**
Total	1,003,401	100.0	179,699	100.0
Age group				
16–20	70,967	7.1	3430	1.9
21–25	81,519	8.1	4041	2.2
26–30	95,072	9.5	5489	3.1
31–35	103,773	10.3	7254	4.0
36–40	102,952	10.3	8694	4.8
41–45	87,738	8.7	9434	5.2
46–50	81,290	8.1	11,057	6.2
51–55	86,423	8.6	15,266	8.5
56–60	67,150	6.7	16,399	9.1
61–65	56,881	5.7	17,615	9.8
66–70	49,454	4.9	17,640	9.8
71–75	44,254	4.4	18,970	10.6
76–80	37,584	3.7	19,373	10.8
81+	38,344	3.8	25,037	13.9
Sex				
Male	486,929	48.5	81,902	45.6
Female	516,472	51.5	97,797	54.4
Economic activity ages 15–74				
Employee: Part-time	125,477	12.5	11,287	6.3
Employee: Full-time	467,795	46.6	25,480	14.2
Unemployed	24,186	2.4	3429	1.9
Full-time student	25,873	2.6	1002	0.6
Retired	114,357	11.4	39,768	22.1
Student	43,411	4.3	2491	1.4
Looking after home or family	58,352	5.8	7205	4.0
Long-term sick or disabled	35,599	3.6	34,104	19.0
Other	23,691	2.4	6582	3.7
Over 75	84,660	8.4	48,351	26.9

Step 2: Use each sample as a reference sample to combine with the subpopulation microdata and estimate $\tilde{p}_i$ according to Approach B in Section 3.3 using the Chen, et al. (2019) method. Here we also use the key variables: Geography, Age group, Sex, Marital Status, Ethnicity and Economic Activity and the main effects design matrix.

Step 3: Estimate the inverse probability weighted (IPW) estimates $\hat{F}_k$. In addition, use the IPW estimates to calculate the marginal counts for the all two-way interactions on the key variables which will be used for the log-linear modelling in (7).

Step 4: Run the log-linear model described in (7) under the pseudo-maximum likelihood estimation method using the all two-way interactions model with estimates from Step 3 to estimate $\hat{\lambda}_k$.

Step 5: Define $\hat{p}_k = 1/\hat{F}_k$ and estimate $\hat{\phi}_k = \hat{\lambda}_k \hat{p}_k$. Calculate the risk measures in (21) and (22) and compare to the true values known from the population.

4.3 | Results

We first describe the results of the log-linear modelling in Step 2 of Approach A of the application study and the justification for using the all two-way interaction model. We estimate the number of sample uniques that are population uniques τ_1 as shown in (1) under the all two-way interaction model. Averaged over 100 samples, the true value is 196.02 (SE 1.49) and the estimate $\hat{\tau}_1$ is 225.94 (SE 0.98). The B_1 goodness-of-fit criteria in (6) averaged over 100 samples is 0.9165 which is below the critical value of 1.96 showing a good fit of the model.

We therefore use the all two-way interaction model for the log-linear models in both approaches A and B of the application study. We use the main effects model for estimating the probability scores and provide a discussion of the implications of these models in Section 4.4 summarizing the results of the application study. Future work will investigate other types of models and the development of goodness-of-fit criteria for this setting.

Table 2 shows the results of the estimation of the risk measures in (21) and (22) for both Approach A and Approach B in the application study. Figure 2 shows the box plots of the same measures.

From Table 2 and Figure 2, we see that Approach B outperforms Approach A with more accurate estimated risk measures compared to their true values for both τ_1 and τ_1^*. We discuss these

TABLE 2 Average of 100 Iterations with Root MSE (simulation standard errors in parenthesis and the Root MSE provided)

		Estimates	
	True values	**Approach A**	**Approach B**
$\tau_1^* = \sum_k I(F_k = 1, F_k^1 = 1)$	2721	2877.03 (4.74) RMSE 156.10	2638.50 (5.70) RMSE 82.70
$\sum_k I(F_k^1 = 1)$	5613	5613	5613
Proportion τ_1^*	0.485	0.513	0.470
$\tau_1 = \sum_k I(F_k = 1, F_k^1 = 1, f_k = 1)$	55.50 (0.76)	50.50 (0.54) RMSE 5.03	54.58 (0.62) RMSE 1.11
$\sum_k I(F_k^1 = 1, f_k = 1)$	381.72 (1.71)	381.72 (1.70) RMSE 1.70	382.21 (1.69) RMSE 1.76
Proportion τ_1	0.143	0.132	0.143
Alternative measure θ in (14)	0.070 (0.0004)	0.072 (0.001)	

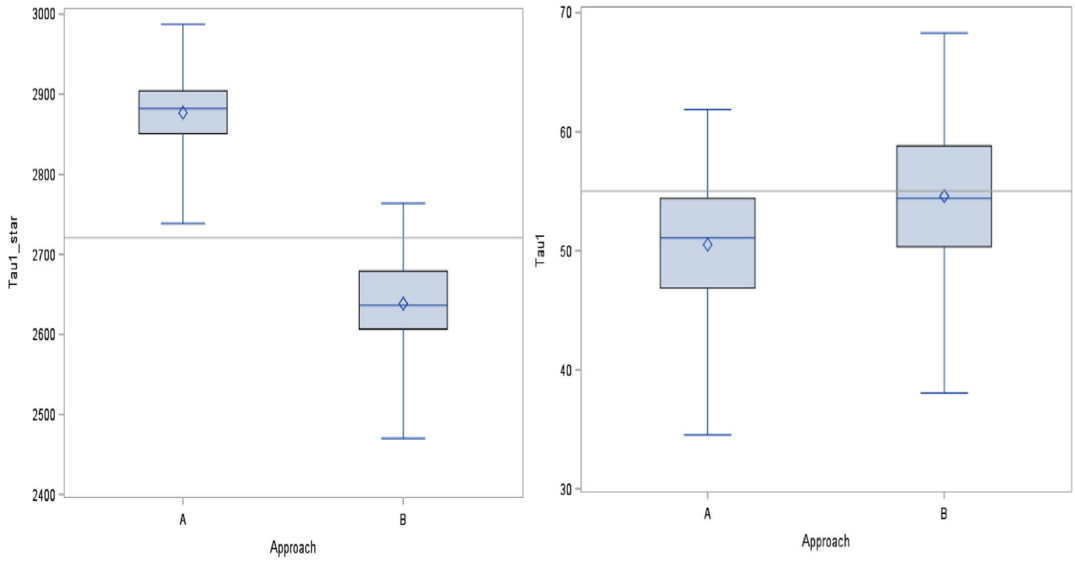

FIGURE 2 Boxplot of τ_1^* (left panel) and τ_1 (right panel) for Approach A and Approach B (the horizontal line is the true value).

findings in Section 4.4. We also see that the risk measure θ in (14) as described in Skinner and Elliot (2002) does not require the use of models and can be estimated without bias. However, the interpretation of this measure may not be as useful compared to the τ_1 measure in (21) for quantifying the risk of re-identification from the perspective of the statistical agency releasing the microdata.

4.4 | Summary of findings

In any two-step estimation approach, the parameters estimated in the second step are dependent on the parameters estimated in the first step. In Approach A, if a parameter λ_k in cell k is estimated as 0 (there is no estimated expected mean (population size) in that cell), this implies that ϕ_k will also be zero because of the relationship that $\phi_k = \lambda_k p_k$. As seen in Skinner and Shlomo (2008) there is monotonicity in the log-linear models to estimate the expected mean parameters based on the observed random sample. The main effects model assumes there are no zeroes in the contingency table defined by the key variables and generally spreads the population mass out too thin, thus lowering the expected mean on the sample unique cells and overestimating the risk of re-identification. On the other hand, the saturated model assumes that all zeros in the contingency table are real zeros and therefore estimates the expected mean to be too high on the sample unique cells, thus lowering the risk of re-identification. The *B*- goodness-of-fit criteria aims to find the right balance in the estimation of the population parameters between the zero cells that are random due to the sampling and the zero cells that are structural (real) zeros. The all two-way interactions model shows a good fit under the log-linear model in Approach A since any zero appearing in an all two-way marginal table is more likely to be a structural zero in the population. Nevertheless, in Approach A there are some cells of the contingency table that have an estimated expected mean equal to zero although there is evidence of population in that

cell because of the presence of individuals in the subpopulation. Thus, we are not utilizing all the information that is available to estimate the disclosure risk measures. As a result of estimating $\hat{\lambda}_k = 0$ in cell k we obtain also that $\hat{\phi}_k = 0$. Therefore, the τ_1^* in Approach A based on the subpopulation uniques is overestimated due to the fact that $\exp(0) = 1$. The risk measure τ_1 depends on both sample uniques and subpopulation uniques and performs better with a smaller bias compared to τ_1^*.

On the other hand, Approach B starts with the larger subpopulation dataset and that data have more information regarding the population zero and unique cells compared to the smaller random sample. We estimate first the $\tilde{p}_i$, $i = 1, \ldots, N_1$ using the main effects model based on the individual units of the subpopulation. Future work will look at the impact of introducing interactions in this model as well. The estimated propensities $\hat{\tilde{p}}_i$ enable the robust estimation of population counts to use in the pseudo-maximum likelihood estimation of the log-linear model in (7). Approach B provides better estimates of both risk measures with a slight downward bias to τ_1^* but an unbiased estimate for τ_1.

Finally, we can demonstrate the approach of Elamir and Skinner (2006) described in Section 2 and estimate the number of subpopulation and population uniques τ_1^* as estimated in (1) with formula in (4) under different log-linear models. Note that in this case, we are not able to carry out a model-search using the $\hat{B}_1$ goodness-of-fit criteria because they are not valid for a non-probability subpopulation. The true value is $\tau_1^* = 2721$ as shown in Table 2. Under the independent log-linear model $\hat{\tau}_1^* = 2783.9$ and under the all two-way interaction model $\hat{\tau}_1^* = 1792.6$. Under a log-linear model with three main effects Geography, Age group and Sex, and three two-way interactions: Marital Status*Ethnicity, Marital Status*Economic Activity and Ethnicity*Economic Activity we obtain $\hat{\tau}_1^* = 2682.8$.

5 | CONCLUSIONS AND FUTURE WORK

The conclusions for assessing disclosure risk in microdata based on the risk of re-identification for subpopulation registers τ_1^* and/or using the subpopulation register to estimate the risk of re-identification in sample microdata τ_1 is to use Approach B. This assumes that the subpopulation is a large enough dataset to allow for more robust estimation of parameters and compensate for the zero cells of the contingency table. An area of future research is to refine the goodness-of-fit criteria for determining both the correct model used in the estimation of the probability scores and the log-linear model under the pseudo-maximum likelihood approach which produce unbiased estimates of the global risk measures. In addition, whereas Rinott and Shlomo (2007b) considered confidence intervals for the global risk measures defined in Section 2, future research is needed to adapt and develop confidence intervals for the global risk measures shown in Section 3.

As seen in the application study in Section 4, a two-step approach to estimate parameters of the disclosure risk measures may not enable estimating the risk of re-identification for samples drawn from the subpopulation, and more generally for non-probability samples. There is less information about the population to compensate for the zero cells in the contingency table due to sampling. Assessing disclosure risk for a non-probability sample is becoming more relevant in recent years with the increased use of non-probability samples to collect data for hard-to-capture populations. For this purpose it is clear that we may need to develop a new approach where the parameters λ_k and p_k are estimated simultaneously using the maximal amount of information from all available sources of data. Future

research will focus on an alternative method. For example, one can treat the problem as estimation with incomplete data and use a fully Bayesian approach or an EM algorithm to estimate the unknown conditional distribution $F_k^1|f_k$ (or $f_k^1|f_k$ assuming a sample from the subpopulation or more generally a non-probability sample) under a complete data likelihood of $\prod_k \Pr(f_k^1, f_k, F_k^1)$.

ACKNOWLEDGEMENT

This research was progressed during the Data Linkage and Anonymization (DLA) Programme of the Isaac Newton Institute for Mathematical Sciences, Cambridge United Kingdom (July-December 2016) EPSRC grant EP/K032208/1.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at https://www.nomisweb.co.uk/census/2001/all_tables

ORCID

Natalie Shlomo https://orcid.org/0000-0003-0701-5080

REFERENCES

Benedetti, R., Capobianchi, A. & Franconi, L. (1998) Individual risk of disclosure using sampling design. Contributi Istat.

Bethlehem, J., Keller, W. & Pannekoek, J. (1990) Disclosure limitation of microdata. *Journal of the American Statistical Association*, 85, 38–45.

Carota, C., Filippone, M., Leombruni, R. & Polettini, S. (2015) Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Annals of Applied Statistics*, 9(1), 525–546.

Chen, Y., Li, P. & Wu, C. (2019) Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115(532), 2011–2021.

Duncan, G. & Lambert, D. (1989) The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, 7, 207–217.

Elamir, E. & Skinner, C.J. (2006) Record-level measures of disclosure risk for survey micro-data. *Journal of Official Statistics*, 22, 525–539.

Fienberg, S.E. & Makov, U.E. (1998) Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics*, 14, 385–397.

Forster, J.J. & Webb, E.L. (2007) Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. *Journal of Royal Statistical Society Series C*, 56, 551–570.

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J. & De Wolf, P.P. (1998) Post randomisation for statistical disclosure control: theory and implementation. *Journal of Official Statistics*, 14, 463–478.

Ichim, D. (2008) Extensions of the re-identification risk measures based on log-linear models. In: Domingo-Ferrer, J. & Saygın, Y. (Eds.) *Privacy in statistical databases*. Lecture Notes in Computer Science 5262. Berlin: Springer, pp. 203–212.

Lambert, D. (1993) Measures of disclosure risk and harm. *Journal of Official Statistics*, 9, 313–331.

Manrique-Vallier, D. & Reiter, J.P. (2012) Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*, 107, 1385–1394.

Paass, G. (1988) Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics*, 6, 487–500.

Rao, J.N.K. & Thomas, D.R. (2003) Analysis of categorical response data from complex surveys: an appraisal and update. In: Chambers, R.L. & Skinner, C.J. (Eds.) *Analysis of survey data*. Chichester, UK: Wiley, pp. 85–108.

Reiter, J.P. (2005) Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100, 1103–1112.

Rinott, Y. & Shlomo, N. (2006) A generalized negative binomial smoothing model for sample disclosure risk estimation. In Domingo-Ferrer, J. & Franconi, L. (Eds.) *Privacy in statistical databases*. Lecture Notes in Computer Science 4302. Berlin: Springer, pp. 82–93.

Rinott, Y. & Shlomo, N. (2007a) A smoothing model for sample disclosure risk estimation. In Liu, R., Strawderman, W. & Zhang, C.-H. (Eds.) *Complex datasets and inverse problems*. Lecture Notes—Monograph Series 54. Beachwood, Ohio: Institute of Mathematical Statistics, pp. 161–171.

Rinott, Y. & Shlomo, N. (2007b) Variances and confidence intervals for sample disclosure risk measures. In: *Bulletin of the International Statistical Institute: Proceedings of the 56th Session of the International Statistical Institute*. ISI'07, Lisbon, pp. 1090–1096.

Shlomo, N. (2010) Releasing microdata: disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality*, 2(1), 73–91.

Shlomo, N. & Skinner, C.J. (2010) Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Annals of Applied Statistics*, 4(3), 1291–1310.

Skinner, C.J. (1992) On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21–32.

Skinner, C.J. & Elliot, M.J. (2002) A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society Series B*, 64, 855–867.

Skinner, C.J. & Holmes, D. (1998) Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 361–372.

Skinner, C.J., Marsh, C., Openshaw, S. & Wymer, C. (1994) Disclosure control for census microdata. *Journal of Official Statistics*, 10, 31–51.

Skinner, C.J. & Shlomo, N. (2008) Assessing identification risk in survey micro-data using log linear models. *Journal of American Statistical Association*, 103(483), 989–1001.

How to cite this article: Shlomo, N. & Skinner, C. (2022) Measuring risk of re-identification in microdata: State-of-the art and new directions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1–19. Available from: https://doi.org/10.1111/rssa.12902