Chris J. Skinner

The probability of identification: applying ideas from forensic statistics to disclosure risk assessment

Article (Accepted version) (Refereed)

The probability of identification: applying ideas from forensic statistics to disclosure risk assessment

C. J. Skinner

University of Southampton, U. K.

Summary. This paper establishes a correspondence between statistical disclosure control and forensic statistics regarding their common use of the concept of 'probability of identification'. The paper then seeks to investigate what lessons for disclosure control can be learnt from the forensic identification literature. The main lesson considered here is that disclosure risk assessment cannot, in general, ignore the search method employed by an intruder seeking to achieve disclosure. The effects of using several search methods are considered. Through consideration of the plausibility of assumptions and 'worst case' approaches, the papers suggests how the impact of search method can be handled. The paper focuses on foundations of disclosure risk assessment, providing some justification for some modelling assumptions underlying some existing record level measures of disclosure risk. The paper illustrates the effects of using different search methods in a numerical example based upon microdata from a sample from the 2001 Census.

Key words: confidentiality; microdata; record linkage; disclosure control; uniqueness.

To appear in *Journal of the Royal Statistical Society, Series A*

1. Introduction

Statistical agencies conducting surveys or censuses need to protect the confidentiality of respondents when releasing outputs (Doyle et al, 2001). A major aim in confidentiality protection is to avoid *identification*. For example, the key 'confidentiality guarantee' in the National Statistics Code of Practice (National Statistics, 2004, p.7) is that 'no statistics will be produced that are likely to identify an individual'. Bethlehem et al. (1990) refer to similar principles elsewhere, such as in the International Statistical Institute Declaration on Professional Ethics. Concern about identification is particularly pronounced for releases of microdata, where the identification of a record in a microdata file might lead to the disclosure of the values of sensitive variables (Paass, 1988; Duncan and Lambert, 1989; Reiter, 2005).

Principles of confidentiality protection, such as that embodied in the National Statistics Code of Practice, are often expressed broadly and require refinement if they are to be implemented in practice. The concept of identification itself seems fairly clear: it involves linking an element of the output, such as a microdata record, with a known individual or other specified unit (Bethlehem et al., 1990). More challenging is the concept of the *probability* of identification, to which confidentiality protection principles often refer. For example, the phrase 'likely to' in the National Statistics confidentiality guarantee is a probabilistic notion. The probability of identification is often referred to as *identification risk* or the *risk of identity disclosure* in the statistical disclosure control (SDC hereafter) literature (e.g. Paass, 1988; Duncan and Lambert, 1989; Reiter, 2005). The assessment of this probability is not straightforward, in particular since the underlying uncertainty might arise from a variety of sources, such as: whether an attempt

at identification by an intruder might take place, what auxiliary information an intruder might be able to use to attempt identification or which elements of the output or known individuals might be selected for an attempt at identification. Some of these sources of uncertainty may be handled by appropriate definition and alternative assumptions, such as via the components of risk approach of Marsh et al. (1991). Nevertheless, there remain challenges in assessing the uncertainty, as will become apparent in this paper.

One field of statistical application where there has been rigorous discussion and development of methods for assessing the probability of identification is forensic science (e.g. Dawid, 1994; Balding and Donnelly, 1995). The aim of this paper is, first, to establish a correspondence between the forensic identification literature and that on SDC and then to consider the relevance of some ideas from the former literature to the assessment of identification risk in an SDC context.

One particular implication of the forensic identification literature, upon which we shall focus, is that the probability of identification may depend upon the *search method* used by an intruder to select an element of the output and a known individual in the population for linking. While the SDC literature has acknowledged that intruders might employ different search methods to improve their chances of disclosure (e.g. Duncan and Lambert, 1989; Lambert, 1993), expressions for identification risk appearing in the SDC literature (e.g. Paass, 1988) are generally not dependent on the search method, for given auxiliary information. Following the forensic identification literature, we shall show how such dependence can arise. This finding makes the task of disclosure risk assessment harder, since the search method employed by a hypothetical intruder is necessarily unknown. We shall discuss how this problem might be addressed.

We shall argue that the assessment of identification risk in SDC may be viewed as a generalization of a forensic identification problem. As a consequence, we shall consider how forensic identification approaches may be extended to identification risk assessment in SDC. Our focus will be on the foundations of risk assessment methodology. We shall, however, outline an application in section 6 and provide some numerical illustrations using data from the 2001 Census.

Our focus in an SDC context will be on microdata, although much of this paper will also be relevant to any form of output where identification is relevant, i.e. where there is concern about the linking of elements of the output to known individuals (or other specified units). Our discussion will apply to cases where SDC methods, such as perturbation (Willenborg and de Waal, 2001), have been applied, provided that each record of the resulting microdata (or element of the output) can still be interpreted as having originated from a given individual. Otherwise, it is not clear that there is reason to be concerned about identification.

We are not the first to observe the connection between forensic science and SDC. The reference to 'fingerprinting' in Willenborg and de Waal (2001) provides a simple example. A deeper but more indirect connection may be traced via discussions of connections between forensic science and record linkage, e.g. Copas and Hilton (1990), and connections between record linkage and SDC, e.g. Paass (1988).

We shall begin in Section 2 by introducing a basic mapping between the two problems of forensic identification and disclosure risk assessment. A formal framework will then be set out in Section 3 to encompass both problems, and it will be indicated how the latter may be treated as a generalization of the former. In Section 4, we restrict

attention to situations where an intruder seeks to achieve identification by a matching approach. The nature of identification risk for this approach and, in particular, the impact of different kinds of search methods are discussed in Section 5, with an illustration in Section 6. Finally, in Section 7 we discuss the broad conclusions and their SDC context.

2. The basic correspondence between forensic identification and SDC

To introduce the correspondence, we first set out the two problems in prototypical form.

In forensic identification (e.g. Balding and Donnelly, 1995), a crime has been committed by an unknown *culprit*, who belongs to a specified *population*. The prosecuting authority identifies a member of this population as a *suspect* and brings the suspect to court. Identification occurs if the suspect and the culprit are identical, i.e. the suspect committed the crime or, in other words, the suspect is guilty. Data relevant to identification consist of values of variables observed both on the suspect and at the scene of the crime, e.g. from fingerprints, DNA profiles or eye witness testimony.

In identification risk assessment for microdata (e.g. Paass, 1988), a microdata file is to be released, based upon data provided by a sample of responding units from a population in a survey or census. The file consists of *records* for these sample units, each with the values of several variables. An *intruder,* i.e third party, has information about one or more known units in the population and seeks to link one of these with one of the records. Identification occurs if the selected known unit is identical to the responding unit which provided data for the record. Data relevant to identification consist of values of variables which are both recorded in the microdata and available to the intruder for the known units. These are often called *key variables*.

The correspondence between the two problems is summarised in Table 1. The crime corresponds to cooperation by a responding unit in a survey or other form of data collection, normally undertaken under some pledge of confidentiality. (Given most agencies' desire to avoid non-response, the correspondence is ironic!) The culprit corresponds to the responding unit. For simplicity, we shall generally suppose that both the culprit and the respondent are individuals. They each belong to some specified population. The prosecuting authority corresponds to the intruder. The suspect, identified by the prosecuting authority, corresponds to the individual chosen by the intruder for linking to a given record in the microdata. To assess the probability that the suspect is guilty, the court will use evidence which links the suspect to the scene of the crime via some shared characteristics, which correspond to the key variables. Some of the other forms of correspondence in Table 1 will be returned to in Section 3.

In the forensic identification problem there is just one crime, one culprit and one suspect. (Note that if the crime is committed by several individuals jointly then we use the term culprit to denote this cluster of individuals. Likewise, the suspect may consist of a cluster of individuals who are suspected to have committed the crime jointly.) The forensic identification problem therefore corresponds to a special case of the disclosure risk assessment problem, where there is just a single record in the microdata and where the intruder links just one known individual to this record. We thus view the SDC problem as generalizing the forensic identification problem to the case where multiple crimes are committed and there are multiple potential suspects that might be linked to these crimes.

3. Formalisation of the correspondence

We now seek to expand upon and formalise the correspondence introduced in the previous section. We begin in Section 3.1. by setting out our general framework for assessing identification risk in the context of SDC. Then, in Section 3.2., we discuss how the forensic identification problem may be considered in this framework.

3.1. SDC problem

We consider a rectangular microdata file in which each record contains values on a common set of variables for a unit in a finite population U of size N. The units might in principle take different forms, for example households or businesses, but here we shall assume that they are individuals for simplicity. The microdata file might have been subject to perturbation by SDC methods, provided that it remains meaningful to associate each record with a unique individual.

We follow Paass (1988) and assume, hypothetically, that an *intruder* seeks to *link* one or more microdata records to one or more *known individuals* in the population using the values of certain *key variables* observed in both the microdata and on the known individuals. The known individuals might be drawn from a different source available to third parties, for example a database consisting of multiple records containing values of the key variables.

We define *identification risk* as the probability that a link between a particular record and a particular known individual is correct, conditional on an intruder having selected this record and this individual for linkage using a specified search method and specified auxiliary information. If the intruder attempts multiple links between several records and several known individuals then there is an identification risk for each

attempted link. Our definition implies a risk for each (record, known individual) pair which might have resulted from an intruder attack and, in particular, for the case when the known individual is in fact the individual to which the record belongs. We shall take the latter case to define the identification risk for a given record. The possible combination of such record level measures of risk to form a file level measure will be discussed in Section 7.

Suppose then that the intruder arrives at a potential link between a microdata record *r* and a known individual in the population, denoted *B*, as a result of using a particular *search method.* The intruder might, for example, begin with a given *target individual, B*, in the population for which additional information is sought, and then search for the record in the microdata which appears to provide the best match to *B*. Let *A(r)* denote the individual to which microdata record, *r,* belongs and write *A(r)* as *A* when this is unambiguous. Identification occurs if *A*=*B*. Note that, in our notation, *A* and *B* represent unique identifiers of units in the population, e.g. names and addresses, whereas *r* belongs to the set *s* of microdata records which are labelled arbitrarily, $s = \{1,...,n\}$. The values of the key variables for *r* and *B* are denoted by $X_{A(r)}$ and $\tilde{X}_B$ respectively, where ~ is used to signify that the key variables may be recorded in different ways in the two sources, for example because of measurement error, different definitions or because some SDC method has been applied to the microdata.

The *identification risk*, may then be expressed as:

$$\text{identification risk} = \Pr(A(r) = B \mid X_{microdata}, \tilde{X}_{population}, \text{search method}), \qquad (1)$$

where $X_{microdata}$ and $\tilde{X}_{population}$ consist of the values assumed available to the intruder on X for records in the microdata and on $\tilde{X}$ for individuals in the population, respectively.

We suppose that the probability in (1) refers to two possible kinds of stochastic mechanism: first, a *superpopulation model* for the generation of the values X and $\tilde{X}$, which may include a stochastic SDC mechanism used to perturb X or measurement error mechanisms affecting both X and $\tilde{X}$; and second, the selection of *r* and *B*, i.e. the combination of the search method and any probability sampling scheme (and nonresponse mechanism) which led to the selection of the respondents, underlying the microdata, from the population.

We may compare the identification risk in (1) with the probability $\Pr(A(r) = B \mid X_{microdata}, \tilde{X}_{population})$, representing the uncertainty faced by the intruder when assessing whether an arbitrary record *r* belongs to an arbitrary known individual *B*, *prior* to any search, assuming the same information on X and $\tilde{X}$ is available. Such probabilities are considered by Paass (1988) and Reiter (2005). If this probability and the probability in (1) are the same then the search method is said to be *ignorable*. If this condition holds then disclosure risk assessment should be easier, since the search method of a hypothetical intruder is necessarily unknown. However, we shall show in section 5.2. that search methods need not necessarily be ignorable and we shall discuss in section 5.4. how we might deal with this possibility.

The probability in (1) is to be interpreted from the perspective of the releasing agency or disclosure auditor, based upon a set of stated assumptions about what auxiliary information might be available and the various stochastic mechanisms above. These assumptions are taken to be ones that could be publicly defended as realistic or correspond to confidentiality protection guidelines.

3.2. Forensic Identification

We now outline the corresponding set-up in forensic identification, following the analogy set out in Section 2. The microdata sample is reduced to a single record *r* corresponding to the *culprit* $A(r)$ committing the crime and *B* becomes the *suspect*, observed to have a particular combination of traits, i.e. key variables, known to be shared by the criminal. For simplicity, we conceive of the criminal and the suspect as individuals, although these might be groups of individuals working together. The population consists of the set of individuals who could have committed the crime and the search method refers to the selection of *B* from this population. There is only one culprit and hence the search method does not refer to the selection of *A*. In the SDC set-up, $A(r)$ might therefore be interpreted as having committed the 'crime' of acting as a respondent in a survey, providing data upon which the given microdata record has been based.

The evidence recovered from the crime scene about the culprit is denoted X_A. The corresponding characteristics of the suspect are denoted $\tilde{X}_B$. Again the key variables may be recorded in different ways, for example if X_A includes variables obtained from eye-witness accounts then these may be subject to measurement error. The identification risk corresponds to the probability that the suspect is *guilty*, that is that *B* is the same person as *A*.

Explicit expressions for this probability of guilt may be obtained under distributional assumptions. For example, for the case where X_A and $\tilde{X}_B$ are normally distributed, Lindley (1977) provides expressions for the likelihood ratio (for $A = B$ vs. $A \neq B$) corresponding to this 'posterior' probability of guilt given the observed values

of X_A and $\tilde{X}_B$. Fuller (1993) provides expressions which may be interpreted as extensions of Lindley's results to the SDC case. Expressions for a further special case will be considered in the next section.

4. Linkage by matching

The discussion in Sections 2 and 3 applies to a very wide class of possible search methods. In practice, an important class of methods, relevant to both SDC and forensic identification, may be defined in terms of matching. In this case, there is a decision rule with a binary outcome, *match* or *non-match,* for any pair $(X_A, \tilde{X}_B)$. Thus, for a given record, *r*, in the microdata with key variable values $X_{A(r)}$ (or analogously a given crime with evidence $X_{A(r)}$ about the culprit), the decision rule defines a set S_r of possible individuals in the population with values of $\tilde{X}_B$ which match $X_{A(r)}$ (and all remaining individuals will not match).

Some examples of how such a matching rule might arise are:

(i) if the key variables are categorical, misclassification is ignored and X_A is said to match $\tilde{X}_B$ if all of the key variables take the same value;

(ii) if the key variables are continuous or categorical and X_A is said to match $\tilde{X}_B$ if measurement error is judged to make X_A and $\tilde{X}_B$ 'indistinguishable' (Balding and Donnelly, 1995, p.36);

(iii) if the key variables are continuous or categorical and a record linkage decision rule of the Fellegi and Sunter (1969) type is used, generating three possible outcomes: 'link', 'non-link' or a 'possible link' for each

pair ($X_A, \tilde{X}_B$). We suppose the 'possible link' category is pooled with one of the other two categories.

Such matching approaches have been widely considered in the forensic identification literature. For example, Kingston (1965) defines identification in terms of the same kind of set S_r as above. We shall return to example (i) in Section 6.

5. Identification Risk for Linkage by Matching

In this section we consider the nature of the probability of identification in (1) for the kinds of linkage methods described in Section 4. Sections 5.1. and 5.2. will focus on the case of a single record, as in forensic identification. The more general case will be considered in Sections 5.3 and 5.4.

5.1. Basic Formulation for a Single Record

We begin by considering an arbitrary record *r* in the microdata, ignoring the remaining microdata records, as in the forensic identification case. We define S_r as in Section 4 and let F_r denote the size of this set. We assume that any discrepancies of measurement between X and $\tilde{X}$ are allowed for in the matching rule sufficiently so that $\tilde{X}_{A(r)}$ matches $X_{A(r)}$, i.e. $A(r) \in S_r$, and thus $F_r \geq 1$.

Suppose that, using the linkage approach, an intruder finds an individual *B* in S_r. We initially assume that F_r is known. By assumption about the linkage rule and the fact that the remaining records are being ignored, the key variable values $X_{A(r)}, \tilde{X}_B$ carry no information about identification, i.e. whether $A(r) = B$, beyond the following

information: $A(r) \in S_r, B \in S_r$ and F_r. Thus, the identification risk in (1) may be expressed as:

$$\text{identification risk} = \Pr(A(r) = B \mid X_{microdata}, \tilde{X}_{population}, \text{search method})$$

$$= \Pr(A(r) = B \mid A(r) \in S_r, B \in S_r, F_r, \text{search method}). \quad (2)$$

Under fairly weak conditions on the mechanism leading to the selection of *r* and *B*, the expression in (2) reduces to

$$\text{identification risk} = 1/F_r \ . \quad (3)$$

For example, (3) holds if the intruder is equally likely to select *B* as any member of S_r, conditional on *r* and the event $B \in S_r$. Assumptions for (3) to hold are also made and justified by Dawid (1994, assumption A1) and Balding and Donnelly (1995, Assumption 1 and equation 7) in the forensic identification context. One circumstance where (3) might be questionable in an SDC context is where the intruder begins with an arbitrary target individual in the population, unequal probability sampling is employed in the selection of the microdata sample and a match is observed which is unique in the microdata. In this case, the F_r possible samples that could lead to this observed outcome are not necessarily equally likely if the probability function in (2) is defined in terms of the sampling scheme. Hence (3) may not hold. Nevertheless, in this case it appears difficult to arrive at an alternative to $1/F_r$ for the right hand side of (3), which is a function of information which an intruder might realistically have in practice, and we shall not pursue such concerns here. For the remainder of the paper we shall suppose that expression (2) does reduce to expression (3).

The simple expression $1/F_r$ in (3) has been noted by several authors, both in the forensic identification literature, e.g. Kingston (1965), and in the SDC literature, e.g. Duncan and Lambert (1989). The difficulty with (3) in practice is that F_r will generally be unknown. Indeed, in the SDC context a key consideration is to ensure that the form of release should not permit key variables to be available where F_r might be known to a potential intruder and be small, say one or two. When F_r is unknown, we remove it from the conditioning set in (2) to give:

$$\begin{aligned}\text{identification risk} &= \Pr(A(r)=B \mid A(r)\in S_r, B\in S_r, \text{ search method})\\ &= \sum_{F=1}^{N} \Pr(A(r)=B \mid A(r)\in S_r, B\in S_r, F_r=F, \text{ search method})\\ &\qquad \times \Pr(F_r=F \mid A(r)\in S_r, B\in S_r, \text{ search method})\\ &= \sum_{F=1}^{N} (1/F)\ \Pr(F_r=F \mid A(r)\in S_r, B\in S_r, \text{ search method}),\end{aligned}$$

under our assumption that (2) reduces to (3), and hence

$$\text{identification risk} = E(1/F_r \mid A(r)\in S_r, B\in S_r, \text{ search method}), \qquad (4)$$

where the expectation is with respect to the conditional distribution $\Pr(F_r \mid A(r)\in S_r, B\in S_r, \text{search method})$ of F_r given the observed events. The problem of determining the identification risk then reduces to one of determining this distribution. We now consider how to obtain an expression for this distribution, following the approach used in the forensic identification literature. This involves specifying both a *superpopulation model*, governing the probability process underlying the event $B\in S_r$, and a search method. Treating the record *r* as fixed, we may specify the superpopulation model by specifying the distribution of the binary indicator variables Z_{ri} for whether $\tilde{X}_i$ matches $X_{A(r)}$ (for individuals *i* in the population). The event $B\in S_r$ then corresponds to

the event $Z_{rB} = 1$ and the assumption that $\tilde{X}_{A(r)}$ matches $X_{A(r)}$ corresponds to the event that $Z_{rA} = 1$. A standard superpopulation model (e.g. Kingston,1965; Dawid, 1994) which treats the size, *N*, of the population as fixed, is that the $Z_{ri}, i \in U$, are independent and identically distributed Bernoulli trials with *p* denoting the probability of a match. This implies that F_r is Binomially distributed with parameters *N* and *p* and we refer to this as the *Binomial model*. The relation between these models and some models used in SDC will be considered in Section 6. We shall treat *p* as known, for simplicity, in the remainder of this section. In forensic identification applications, *p* will often be estimated from a population database, possibly one from which a suspect has been selected. In SDC applications, *p* might similarly be estimated from a database available to an intruder, but also from multiple records in the microdata, as will be discussed in Section 6. The latter option has no analogue in forensic identification.

In the following section, we set out a number of possible search methods considered in the forensic identification literature and discuss the nature of the conditional distribution for F_r and the expression for the risk in (4) given these search methods and the Binomial model.

5.2 Search Methods from forensic identification literature

In this section, we describe a series of search methods, labelled *r1, r2,…* to signify that the search begins with a specified *record*.

Search Method r1: *suspect is selected by searching the population randomly until a match is found.*

This method may be illustrated in the SDC context by the 'journalist scenario' of Paass (1988), where a journalist selects a record from the microdata with an unusual

combination of values of the key variables and tries to find an known matching individual in the population by searching through sources accessible to the journalist until a match is found. The implicit assumption here is that the 'systematic' element of the journalist's method of searching the population is fully captured by the matching rule and that, otherwise, the search is equally likely to lead to any one of the F_r members of S_r.

Under this search method, we may write $\Pr(F_r \mid A(r) \in S_r, B \in S_r$, search method) $= \Pr(F_r \mid A(r) \in S_r$, search method) since, conditional on $A(r) \in S_r$, the event $B \in S_r$ is not informative about F_r because some match must be found if we search long enough. The event $A(r) \in S_r$ tells us that $Z_{rA} = 1$ but, under the Binomial model, is not informative about Z_{ri} for $i \neq A$ and so the conditional distribution of F_r is obtained by writing $F_r = 1 + (F_r - 1)$ and noting that the conditional distribution of $F_r - 1$ given $A(r) \in S_r$ under this search method is Binomial with parameters $N-1$ and p (Lenth, 1986; Dawid, 1994, p.167). Straightforward calculation using the Binomial density shows that the expectation in (4) has the closed form expression:

$$\text{identification risk} = [1-(1-p)^N]/[Np]. \qquad (5)$$

An implicit assumption here is that *N* and *p* are known. A further assumption is that *y*, the number of non-matches arising before the intruder finds a match, is unrecorded and hence not conditioned upon. The effect of recording *y* will be considered in method *r3*.

Search Method r2: *suspect is drawn at random from the population and found to match.*

This method appears less plausible in the SDC context, since the expected payoff to a potential intruder seems likely to be too low if no search is undertaken. The nearest parallel appears to be the case of 'spontaneous recognition' (Willenborg and de Waal,

2001, p.62) where an intruder happens, by chance, to observe a match between a microdata record and a known individual.

For this search method, the event $B \in S_r$ is informative about F_r, making larger values of F_r more likely. We may write:

$$\Pr(F_r \mid A(r) \in S_r, B \in S_r) \propto \Pr(F_r \mid A(r) \in S_r) \Pr(B \in S_r \mid F_r, A(r) \in S_r), \qquad (6)$$

where implicitly each term also conditions on the search method. The first term on the right hand side of (6) is the density function of $F_r \sim 1 + Bin(N-1, p)$, as for method *r1*. The second term equals F_r / N since we assume the suspect is drawn randomly. We may interpret the implied distribution $\Pr(F_r \mid A(r) \in S_r, B \in S_r)$ as a 'size-biased' Binomial distribution (Dawid, 1994; Balding and Donnelly, 1995). It is straightforward to show that the constant of proportionality in (6) is $N/[1+(N-1)p]$ and hence that the conditional expectation in (4) takes the form:

$$\text{identification risk} = 1/\{1+(N-1)p\}. \qquad (7)$$

Search Method r3: as search method r1 but where the length of the search is recorded.

If *y* is recorded then the event $B \in S_r$ does become informative about F_r, as for method *r2*. Indeed, if $y = 0$, methods *r2* and *r3* are identical. To obtain the conditional distribution of F_r of interest, all components of expression (6) may be modified by including the event of *y* previous non-matches alongside the conditioning event $A(r) \in S_r$. This simply has the effect of replacing N by $N - y$ in each of the terms on the right hand side of (6) and hence (c.f. Dawid,1994; Balding and Donnelly, 1995) expression (7) is modified to:

$$\text{identification risk} = 1/\{1+(N-1-y)p\}. \qquad (8)$$

Search Method r4: suspect is found to be unique match in a database.

If a search is made among $y+1$ potential suspects in a database, the same probability calculations may be made as for method *r3* with y known (Balding and Donnelly, 1995) and so the identification risk is the same for these two methods. In the SDC context, this method corresponds again to the journalist scenario where the database represents a particular source available to the journalist.

These results for methods *r3* and *r4* have been subject to some debate in the forensic identification literature. Expression (8) implies that the greater the value of y, i.e. the longer the search, the greater the risk of identification, although this increase will tend to be minor if the fraction of the population searched, y / N , is small. This contrasts with an alternative argument, advanced for example by Stockmarr (1999), that the risk may be severely reduced by such a database search. See Dawid (2001) and Balding (2002) for some of the ensuing debate. To illustrate this debate in an SDC context, suppose that a journalist claims to have found a unique match between a named individual and a record in a public use microdata file released by a statistical agency. On discussion, the journalist admits to have found the match by searching through a large database of 100,000 individuals. The agency might claim, following the alternative argument, that it is not surprising that a match has been found as a result of such an extensive search and argues that, as a result, little weight should be given to the observed match, i.e. the probability that the match is correct should be treated as small. This paper's position, following e.g. Balding (2002), is to suggest that such an argument would be misleading. It is true that the probability of finding a match does increase the longer the search and thus that the journalist's discovery may be unremarkable overall. Nevertheless, for the

particular record for which a match is found, the fact that a further proportion of the population has been searched for a match without success increases rather than decreases the probability that the match is correct. The issue is then whether the value of this increased probability for this record (i.e. expression (8) under the Binomial model, assuming p is known) is of concern.

Search Method r5: method r1 is extended by continued searching.

If the search is continued without a further match being found then this method may be treated as equivalent to either methods *r3* or *r4*, with y equal to the number of non-matches. If the continued search leads to another individual being found which matches, then Dawid (1994) provides an expression for the resulting risk, assuming y is not recorded. In the extreme, if a complete search of the population revealed F_r, the number of individuals in the population matching A, the risk would again become $1/F_r$, as in (3).

5.3. Generalization: Search Methods for SDC

Attention was restricted to the case of a single record in the previous two sections. In the general SDC setting, however, there will be multiple records in the microdata. Possible extensions of the previous search methods to this case will be considered in this section and are summarised in Table 2. These extensions are of two types, termed *fishing* and *directed searches* by Paass (1988).

In a fishing method, the intruder first selects a record (or records) in the microdata, possibly a record that he/she expects to be easier to identify as a result of having unusual values or combinations of values of key variables. For example, Paass (1988) considers an expenditure survey, where an intruder might select an individual purchasing two or more boats. The intruder then seeks to find a match for this record using one of the

methods *r1, r2*... above. The judgement about the record being unusual might be based upon the microdata, in the extreme if the record is unique in the sample with respect to *X* (i.e. does not match any other record). We let *r1u*, *r2u*… denote the use of search methods *r1*, *r2*… for a record which is unique in this sense. We treat the case where the intruder selects multiple records for linkage as repetition of methods *r1*, *r1u* etc.

In what Paass (1988) refers to as a directed search, the intruder begins with a known target individual (or individuals) in the population and then searches for a match in the microdata. Out of six scenarios considered by Paass (1988), only one (the journalist scenario above) involves fishing. The remaining five are directed searches. In three of these, it is assumed that the intruder begins with a particular individual in the population and then searches the microdata file for a match. In the remaining two scenarios the intruder begins with a set of known individuals in the population and then seeks matches for each of these in the microdata file. Duncan and Lambert (1989), Lambert (1993) and Reiter (2005) also focus on the case of a directed search.

By interchanging the role of the known population individuals and the microdata records, the search methods in Section 5.2 may be transposed to the case of a directed search. We assume that any intruder who has managed to gain access to the microdata would search the whole file and would not stop at an intermediate stage, for example, at the first match to the target individual, *B*. We thus reject the counterparts of methods *r1*, *r2* and *r3* as unrealistic, since they involve either stopping (*r1* and *r3*) or no search at all (*r2*). The counterpart of method *r4*, treating the microdata file as the counterpart of the database, is:

Search method B1: for a given target individual B, a unique matching microdata record is found.

The use of *B* in the notation *B1* is intended to signify that the search begins with a specified known individual *B*. The intruder cannot search for matches among individuals falling outside the microdata sample and thus the counterpart of method *r5* is rejected as impossible. We also reject methods which generate more than one match in a search of the microdata, on the grounds of restricting attention to worst cases. It would be possible to qualify method *B1* by some method for selecting the target individual. For example, a method which selected the individual as unique within a database might be denoted *B1u*. It would also be possible for the intruder to select more than one known individual for linkage, for example the set of individuals within a database, resulting in an effective repetition of method *B1*. We shall, however, only explore such extensions implicitly through consideration of *B1*.

5.4. Generalization: Risk assessment for SDC

In this section, we consider the generalization of the results on identification risk in Sections 5.1 and 5.2. to the case of SDC for the search methods discussed in Section 5.3. We also seek to compare these methods with respect to risk in order to narrow the class of search methods which it is reasonable for a disclosure risk assessor to consider. This is desirable in practice since dependence of the risk upon the search method complicates the task of the assessor, given that the intruder will generally be hypothetical and hence the search method unknown. We shall argue in this section that it is reasonable for the assessor to restrict attention among the search methods to *r1u* and *B1* and their extensions

to repeated records or known individuals. We consider two types of search methods in turn, under the headings discussed in the Section 5.3.

5.4.1. Fishing methods

We suppose first that the intruder begins by selecting a microdata record and then seeks a match in the population. The expressions for identification risk in Section 5.2. were derived for arbitrary records and hence will still apply provided the selection does not depend on some event which is informative about F_r and any information provided by other records is ignored. Consider, following an example of Paass (1988), the case of an expenditure survey where there is a separate code in the microdata for individuals who purchase two or more boats. In one form of attack, an intruder might decide in advance to select any individual who falls into this category for a matching attempt on grounds of prior judgement that this is an unusual category. In this case, this selection is not dependent on any observed event and the expressions for identification risk in Section 5.2. will continue to apply, under the assumptions made there provided we ignore observed data from other records. (This argument might be formalised under a given superpopulation model using the irrelevance of stopping rules, following Berger and Wolpert, 1984, p.74).

In a second form of attack, the intruder might seek an unusual category on the basis of observing the microdata, for example it might be observed that there is only one individual in the microdata who purchases two or more boats. Here, conditioning the risk on the search method (see (1)) corresponds to conditioning on this observed sample uniqueness. These two forms of attack correspond to the distinction between methods *r1*, *r2*,… of Section 5.2. and methods *r1u*, *r2u*,…of Section 5.3.

It follows from our definition in (1), however, that even in the case of method *r1* we should condition the risk on $X_{microdata}$, i.e. the information provided by other microdata records and, in particular, sample uniqueness if it occurs. Thus the risk for an individual in the microdata who was selected by method *r1* and subsequently found to be sample unique should be the same as if the same individual was selected by the intruder using method *rlu* after observing sample uniqueness. The risk for method *r1* will tend to be higher if it is observed that the individual is not sample unique and hence, if concern is with the worst cases, we may argue that it is sufficient to restrict attention to *r1u*.

In fact, if the sampling fraction is small, as is common in many SDC applications, sample uniqueness will not carry much information about F_r under the Binomial model where *p* is given, since F_r will be primarily determined by the behaviour of non-sample individuals. See section 6 for more detailed discussion of this point. We may therefore expect the risk for methods *r1u*, *r2u*,… to be very similar to that for methods *r1*, *r2*,… in these circumstances. For simplicity, we shall now compare risk for the latter methods and then infer that similar comparative properties will apply to the former methods. We suppose that the event of sample uniqueness represents the worst case, in terms of what microdata information the intruder might use to select a record for matching, and thus suppose that it is unnecessary to consider conditioning (1) on other features of $X_{microdata}$.

Suppose then that one of the methods *r1, r2, … , r5* is employed and that the selection of the record is not informative so that the expressions for identification risk in Section 5.2. still apply. Note that these results also depend upon assumptions about the sampling scheme, discussed in Section 5.1. We now consider each of methods *r2, … , r5*

in turn, comparing them with method *r1*, and argue that it is reasonable to restrict attention amongst the search methods *r1* to *r5* just to method *r1*.

Consider first search method *r2*. We have already suggested in Section 5.2. that this method is less plausible than the other methods. Moreover, search method *r1* will, in general, lead to higher risk than method *r2* since the size biasing in the latter method makes larger values of F_r more likely and these are associated with lower risk. Balding and Donnelly (1995) give an example where N=101, p=0.004 and the expressions in (5) and (7) are 0.826 and 0.714 respectively. Thus, disregarding method *r2* but considering method *r1* will be a conservative approach to risk assessment.

Methods *r3* and *r4* may lead to slightly higher risk than method *r2*, but the risk will only be higher than method *r1* if a substantial proportion of the population is searched. For example, if N=101 and p=0.004 then $1/\{1+(N-1-y)p\} > 0.826$ requires $y \geq 48$, i.e. almost half the population must be searched. Indeed, using the approximation that N is large, p is small and $Np << 1$ considered in Balding and Donnelly, it will in general be necessary for at least half the population to be searched (i.e. $y/N > 0.5$) for methods *r3* or *r4* to lead to a higher risk than method *r1*. Principles governing SDC often enable such 'disproportionate' amounts of intruder information to be ruled out. For example the National Statistics Code of Practice (National Statistics, 2004, pp.7, 8) states, in relation to the use of SDC methods, that assumptions about the "information likely to be available to third parties" should be made "against the following standard: it would take a disproportionate amount of time, effort and expertise for an intruder to identify a statistical unit to others, or to reveal information about that unit not already in the public domain".

Method *r5* may also be disregarded on the grounds that the only relevant cases under this method reduce to those under methods *r3* and *r4* since it seems reasonable to discount the possibility of the intruder reporting that they have found a second match, because this would be expected to substantially reduce the risk by ruling out the possibility of population uniqueness. The resulting risk would be at most 0.5. For example, for the case N=101, p=0.004, expression (4.8) in Dawid (1994) implies the risk is 0.467.

Finally, let us turn to methods *r1u-r5u.* As discussed earlier, we may expect the risk for these methods to be similar to that for methods *r1-r5* for a given selected individual and thus we suggest the above argument for restricting attention to *r1* may be extended to justify restricting attention to *r1u* out of the former methods. As noted earlier, it is appropriate to condition the risk for *r1* on the observed occurrence or otherwise of sample uniqueness and, taking the worst case, since the risk of *r1* given sample uniqueness is the same as the risk for *r1u*, we argue it is sufficient to restrict attention to the latter method.

5.4.2. Directed Searches

Turning to method *B1*, we note first that it is isomorphic to method *r4* if we interchange the role of the microdata and the database. Under this isomorphism, the indicator variables Z_{ri} are translated into variables $\tilde{Z}_{Bi}$ for individuals $i \in U$, indicating whether X_i matches $\tilde{X}_B$. For individuals *i* outside the microdata sample, X_i is defined to contain the values of the key variables which would be recorded in the microdata if *i* were selected into the sample. It is assumed that $\tilde{Z}_{BB} = 1$. The corresponding Binomial model is that the $\tilde{Z}_{Bi}$ are independent and identically distributed Bernoulli trials with $\tilde{p}$ denoting the probability of a match.

It then follows, as above, that under this new Binomial model, the identification risk is given by

$$\text{identification risk} = 1/\{1+(N-n)\tilde{p}\}, \tag{9}$$

where n is size of the microdata sample.

We expect that, for the same individual, p and $\tilde{p}$ will be of similar magnitude in many practical applications. As discussed in the previous section, expression (9) (with $\tilde{p} = p$) will only be greater than the risk for method *r1* if the sampling fraction, n/N, is high, roughly greater than 0.5. Since we expect the risks for *r1u* and r1 to be similar, we expect that in cases with small sampling fractions, it will usually be reasonable for the disclosure risk assessor to disregard *B1* in favour of *r1u*.

6. An Application with Categorical Key Variables and No Misclassification

We now illustrate the assessment of risk in one kind of SDC application which arises with sample microdata from population censuses or social surveys. It is assumed that the key variables are categorical and identically measured in the two sources, with linkage based upon exact matching, i.e. example (i) of section 4. In this case, we label the combinations of categories of the key variables by x so that the earlier expression X for the key variables may now take the integer values $x = 1,...,K$. These combinations may be interpreted as cells in a multi-way contingency table. The Binomial model considered earlier implies a multinomial model for this contingency table. Since we assume that $\tilde{X}$ is identical to X and that linkage is based upon exact matching, the Binomial model in section 5.1 for a given record with $X = x$ implies that the events $X_i = x$ for different population units $i \in U$ are independent and identically distributed with $\Pr(X_i = x) = p_x$,

where the subscript x is added to the probability p to indicate that this model relates to the event $X_i = x$. Assuming that the Binomial model holds for all records with all possible values $x = 1,...,K$, it follows that the X_i are independent and identically distributed random variables with $\Pr(X_i = x) = p_x$, $x = 1,...,K$, $\sum p_x = 1$. $x = 1,...,K$. (Since $\tilde{X} = X$ and hence $\tilde{p} = p$, this model is also a consequence of the Binomial model in section 5.4.2.) The population counts F_x in the cells x thus follow a multinomial distribution with parameters p_x and N, $x = 1,...,K$. A related model, more common in the SDC literature, is the Poisson model where the F_x are independently distributed as $F_x \sim Poisson(\lambda_x)$. The multinomial model can, in fact, be derived from the Poisson model by conditioning on $N = \sum F_x$ and setting $p_x = \lambda_x / \sum \lambda_x$ (McCullagh and Nelder, 1989, p.165). Even unconditionally, it may be argued that the two models have very similar SDC consequences when the p_x are small and N is large (Chen and Keller-McNulty, 1998).

In practice, the p_x are unknown, but inference about them may be made using the multiple records of the microdata. As discussed in section 5.1., we may suppose that an intruder could not know the values of the F_x but he/she may be expected to be able to compute the corresponding sample counts f_x from the multiple microdata records. In typical SDC applications, interest will focus on the 'riskiest' cells where f_x is small, say one or two (the values of p_x for empty cells with $f_x = 0$ will not be of interest since these cells contain no microdata records susceptible to identification). The data within a cell x with such a small value of f_x will, however, carry little information, on its own,

about p_x. For the model to be useful for risk assessment, it is therefore natural to consider 'borrowing information' between cells by modelling the relation between the p_x in different cells. One approach is to consider a compound model, such as a Poisson-gamma model (Bethlehem et al., 1990) where the λ_x, $x = 1,...,K$, are independent and identically gamma distributed or a Dirichlet-multinomial model where the p_x follow a Dirichlet distribution. Such models imply that the identification risk is the same for each microdata record, since they treat all cells as exchangeable and make no use of the key variable characteristics used to construct the cells. Such characteristics may be conditioned upon in a log-linear model, relating p_x or λ_x to main effects and interactions between the key variables (Skinner and Holmes, 1998; Elamir and Skinner, 2006), in order to obtain more 'realistic' probabilities of identification, which may vary across cells. We now illustrate this with a numerical example, drawing on Skinner and Shlomo (2005).

The data come from the 2001 United Kingdom Census for two large areas with a combined size of $N \approx 950,000$ individuals. A simple random sample of size $n \approx 0.005N \approx 4,750$ is drawn from this 'population' to mimic a sample survey. The advantage of using census data is that the population characteristics can be used to validate sample-based procedures.

The following six key variables (with numbers of categories in parentheses) are used: area (2), sex (2), age band (18), marital status (6), ethnicity (17) and economic activity (10). The categories are the same as those used for the Samples of Anonymised Records from the census. See Dale and Elliot (2001) for a discussion of the choice of key variables in similar settings. The number of key variable combinations is thus K = 73,440

= 2×2×18×6×17×10. We assume the multinomial model above, that is that the population counts F_x in the cells of the six-way contingency table formed by cross-classifying the key variables are generated by a multinomial distribution with parameters N and p_x, $x = 1,...,K$. As above, we suppose the F_x are unknown but the corresponding sample counts f_x are known ($\sum F_x = N$, $\sum f_x = n$) and may be used to make inference about the parameters p_x. We suppose that such inference is conducted using a log-linear model for p_x including all main effects and two-way interactions (e.g. Agresti, 2002). Using the population data for validation, this model has been found to generate 'reasonable' disclosure risk assessments both for these data (Skinner and Shlomo, 2005) and similar data sources (Skinner and Holmes, 1998; Elamir and Skinner, 2006). Let $\hat{p}_x$ denote the maximum likelihood estimate of p_x under this multinomial log-linear model. In Table 3 we present values of $\hat{p}_x$ for three individuals selected from the sample. We consider only the 739 sample unique cells, i.e. cells where $f_x = 1$, to continue our 'worst case' analysis, and select those sample unique individuals with the minimum, median and maximum values of $\hat{p}_x$ across these 739 cells. A comparison of the second and third columns in the table shows how the values $\hat{p}_x$ could help the intruder infer which of the sample uniques are likely to have smaller (or larger) values of F_x. For example, individuals in ethnic minority groups tend to fall into cells with smaller values of F_x and this is picked up by the model through the main effect term for the ethnic group. Unusual combinations of pairs of key variables, e.g. widowed 20-24 year-olds, are picked up through the two-way interaction terms in the model. Impossible two-way combinations,

e.g. married 0-4 year olds, can also be handled in the model and will, of course, not appear in the sample.

Table 3 also includes estimates of identification risk for these three individuals under different assumptions about the intruder's search method. Considering first search method *r1* and replacing p by $\hat{p}_x$ in expression (5) gives risk estimates of 0.9298, 0.0149 and 0.009 for the sample unique individuals with minimum, median and maximum values of $\hat{p}_x$ respectively. We might conclude that the release of the sample microdata are not 'likely to identify' the second and third individuals, in the language of the Code of Practice. However, the risk for the first individual appears high. In fact, the first individual is not population unique. There are, in fact, five women in the second area in the population who are recorded as being aged 20-24, of separated marital status, in the Bangladeshi ethnic group and with 'looking after home' as their economic activity. Out of the ten sample unique individuals with the lowest values of $\hat{p}_x$ just three turn out to be population unique so the risk estimate of 0.9298 might be judged somewhat high. This raises questions about the choice of the log-linear model and the estimation method which we shall not pursue here. Our focus is on the comparison of risk estimates for different search methods treating these values of $\hat{p}_x$ as realistic and given.

The above risk estimates for method *r1* only use the microdata to estimate p_x and ignore the information that the individuals are sample unique. As discussed in Section 5.4.1., conditioning on sample uniqueness is equivalent to considering method *r1u*. The microdata sample is obtained by simple random sampling of size n (without replacement) so, under the multinomial model, the conditional distribution of F_x given $f_x = 1$ may be

obtained using the fact that the frequencies, f_x and $F_x - f_x$ are independently Binomially distributed: $f_x \sim Bin(n, p)$ and $F_x - f_x \sim Bin(N-n, p)$. Hence the risk is given by

$$E(1/F_r \mid f_r = 1) = E(1/[1+(F_r - f_r)]) = [1-(1-p)^{N-n+1}]/[(N-n+1)p], \qquad (10)$$

i.e. as in expression (5) but with N replaced by $N-n+1$. Table 3 shows that the impact of this change is minor in all three cases. Inspection of expressions (5) and (10) indicates that this will generally be the case if the sampling fraction *n/N* is small.

We next consider search method *r2*. As expected from the discussion in section 5.4.1, Table 3 displays lower risk estimates for this method than method *r1*, although the reduction is not great. The risk for methods *r3* and *r4* is given by expression (8). This expression depends on the number *y*+1 of individuals in the database used for matching (in method *r4*). We now calculate how large *y*+1 must be for the risk of methods *r3* or *r4* to exceed the risk of method *r1u*. Equating (8) and (10) and solving for *y*+1, we find that it is necessary to search databases of sizes at least 463,000, 17,300 or 5,550 for the minimum, median and maximum cases respectively. Most importantly, we find that for the most risky case we must search a database of almost half the size of the population for method *r4* to lead to a higher risk than method *r1u*. This accords with the discussion in section 5.4.1.

Finally, we consider method *B1*. Expression (9) is the same as expression (8) for method *r4* when $\tilde{p} = p$ (which we are assuming in this section) and when the size *y*+1 of the database is the same as the sample size n. We have $n \approx 4750$ and since this is smaller than the three database sizes above, method *B1* always leads to a lower risk than

method *r1u*. In fact, it gives very similar results to method *r2* since the sampling fraction is small.

In summary, we conclude from the numerical comparison in Table 3 that it is sufficient to consider only method *r1u* as a worse case, but that the values of the risk for all the methods are of a similar magnitude so that the 'worst case' approach would not be overly conservative if an intruder employed one of the other methods. Following the discussion in section 5, we may expect to be able to generalise the finding that method *r1u* is the worst case to any situation where the sampling fraction is small, $X = \tilde{X}$ and any errors in the estimation of p_x can be ignored.

7. Discussion

A key objective of SDC for microdata release is to limit the ability of an intruder to achieve identification. This requires limiting the identification risk for any record which might be selected for linkage. A main theme of this paper has been to consider, following discussion in the forensic identification literature, how this identification risk may depend upon the search method used by the intruder to select the record for linkage. We have discussed how this dependence might occur and have suggested that, in practice, it may be handled by considering worst cases amongst a number of plausible alternative search methods. Our discussion suggests a focus on the method denoted *r1u* earlier, a focus which is consistent with the modelling foundations of approaches in Skinner and Holmes (1998) and Elamir and Skinner (2006) for categorical key variables. This paper therefore provides some justification for the assumptions in these two papers. Another possible analogous application of the ideas in this paper would be to the assessment of the threat

to disclosure posed by record linkage methods based upon a mixture of categorical and continuous key variable.

In this paper we have defined identification risk at the record level. In practice, it will often be necessary in SDC work to make judgements about risk at the file level. This may be achieved by aggregating record level measures, as discussed by Lambert (1993), or by defining a file-level measure directly. For example, Skinner and Elliot (2002) consider two measures which may be represented as two alternative averages of values of $1/F_r$ (c.f. equation (4)). The two measures correspond to alternative possible search methods: one to the intruder drawing a sample unique record in the microdata at random (with equal probabilities); the other to the intruder selecting at random any population unit which match a sample unique record. The two measures can, however, take very different values, illustrating how such file level measures can be very sensitive to assumptions about the intruder's search method. Indeed, we suggest this sensitivity is rather greater than the dependence of record level measures upon the search method, as discussed in this paper.

There are a number of other pros and cons of file-level vs. record level measures. File-level measures, such as the population-averaged measure in Skinner and Elliot (2002), not only have the advantage that they are simple but they can also be estimated robustly, whereas inference about record-level measures may be expected to be more model-sensitive (given the relative amount of 'information' available at each level). On the other hand, the aims of SDC are often expressed in a way that seems to correspond better to a definition at the record level. For example, the requirement in the National Statistics Code of Practice (National Statistics, 2004, p.11) that "the guarantee of

confidentiality must be applied equally to all statistical units" suggests reference to a unit level concept of disclosure risk. Model sensitivity can be handled in the framework of a sensitivity analysis, which seems a necessary feature of disclosure risk assessment in any case if alternative sets of assumptions about the possible auxiliary key information available to an intruder are to be considered, as well as alternative assumptions about possible measurement error. The sensitivity analysis could also be used to handle the selection of alternative types of records. If it is desired to aggregate record level measures across records and if large values of the measure are of most concern then a suitable approach might be count the number of records for which the identification risk is above a given threshold (Lambert, 1993, p.317). Combining record-level measures by counting or averaging seems likely to be more robust to model specification than taking the maximum value.

Acknowledgements

I am grateful to Natalie Shlomo for providing the numerical estimates in Section 6 and Table 3, to the Office for National Statistics for the use of the census data source and to two reviewers for comments.

References

Agresti, A. (2002) *Categorical Data Analysis*. 2nd Ed. New York: Wiley.

Balding, D. J. (2002) The DNA database search controversy. *Biometrics*, **58**, 241-244.

Balding, D. J. and Donnelly, P. (1995) Inference in forensic identification. *J. R. Statist. Soc*. A, **158**, 21-53.

Berger, J.O. and Wolpert, R.L. (1984) *The Likelihood Principle.* Hayward: Institute of Mathematical Statistics.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990) Disclosure control of microdata. *J. Amer. Statist. Ass.,* **85**, 38-45.

Chen, G. and Keller-McNulty,S. (1998) Estimation of identification disclosure risk in microdata, *Journal of Official Statistics*, **14**,79-95.

Copas, J.B. and Hilton, F.J. (1990) Record linkage: statistical models for matching computer records (with discussion). *J. R. Statist. Soc.* A, **153**, 287-320.

Dale, A. and Elliot, M. (2001) Proposals for 2001 samples of anonymized records: an assessment of disclosure risk. *J. R. Statist. Soc.* A, **164**, 427-447.

Dawid, A.P. (1994) The island problem: coherent use of identification evidence. In *Aspects of Uncertainty: A Tribute to D.V.Lindley* (eds P.R. Freeman and A.F.M.Smith) Chichester: Wiley, 159-170.

Dawid, A.P. (2001) Comment on Stockmarr (1999) and author's reply. *Biometrics*, **57**, 976-980.

Doyle, P., Lane, J. I., Theeuwes, J. M. and Zayatz, L.V. eds. (2001) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: North-Holland.

Duncan, G. and Lambert, D. (1989) The risk of disclosure for microdata. *J. Bus. Econ. Statist.*, **7**, 207-217.

Elamir, E.A.H. and Skinner, C. (2006) Record level measures of disclosure risk for survey microdata. *J. Off. Statist.* to appear

Fellegi, I.P. and Sunter, A.B. (1969) A theory for record linkage. *J. Amer. Statist. Ass.,* **64**, 1183-1210

Fuller, W.A. (1993) Masking procedures for microdata disclosure limitation. *J. Off. Statist.,* **9**, 383-406.

Kingston, C.R. (1965) Applications of probability theory in criminalistics. *J. Amer. Statist. Ass.*, **60**, 70-80.

Lambert, D. (1993) Measures of disclosure risk and harm. *J. Off. Statist.,* **9**, 313-331.

Lenth, R.V. (1986) On identification by probability. *J. Forensic Sci. Soc.*, **26**, 197-213.

Lindley, D.V. (1977) A problem in forensic science, *Biometrika*, **64**, 207-213.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991) The case for a sample of anonymized records from the 1991 census. *J. R. Statist. Soc.* A, **154**, 305-340.

McCullagh, P. and Nelder,J.A. (1989) *Generalized Linear Models*, 2nd. Ed. London: Chapman and Hall.

National Statistics (2004) *Code of Practice: Protocol on Data Access and Confidentiality.* Norwich: Her Majesty's Stationary Office.

Paass, G. (1988) Disclosure risk and disclosure avoidance for microdata. *J. Bus. Econ. Statist.,* **6**, 487-500.

Reiter, J.P. (2005) Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Ass.,* **100**, 1103-1112.

Skinner, C.J. and Elliot, M.J. (2002) A measure of disclosure risk for microdata. *J. R. Statist. Soc.* B, **64**, 855-867.

Skinner, C.J. and Holmes, D.J. (1998). Estimating the re-identification risk per record in microdata. *J. Off. Statist*., **14**, 361-372.

Skinner, C. and Shlomo, N. (2005) Assessing disclosure risk in microdata using record level measures. Invited paper, worksession on statistical data confidentiality, United Nations Economic Commission for Europe, Geneva (www.unece.org/stats/documents/2005.11.confidentiality.htm)

Stockmarr, A. (1999) Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. *Biometrics*, **55**, 671-677.

Willenborg, L. and de Waal, T. (2001) *Elements of Statistical Disclosure Control*. New York: Springer.

Table 1. Correspondence between Two Prototypical Problems

Notation	Forensic Identification	Statistical Disclosure Control
	Crime	Responding to a survey
A	*Culprit*, committing crime	*Respondent*
U ($A \in U$)	*Population* of possible culprits	*Population,* from which respondent drawn
	Prosecuting authority bringing suspect to trial	*Intruder*
B ($\in U$)	*Suspect* selected by prosecuting authority	Known individual linked by intruder to microdata record
r	Label for evidence at scene of crime	Label for microdata *record* derived from respondent's data
$A(r)$	Culprit producing evidence r	Respondent providing data in record r
$X_{A(r)}$	Traces of culprit at crime scene	*Key variable values* on record r
$\tilde{X}_B$	Characteristics of suspect corresponding to variables in X	Key variable values observed on individual B
	Search method (selection of B)	Search method (Scenario of attack) (selection of r and B)

Table 2. Alternative Intruder Search Methods

Notation	Starts with selection of:	Proceeds by:
Fishing Methods		
r1	Arbitrary record, *r*	searching population randomly until match is found
r2	Arbitrary record, *r*	drawing individual at random from the population and observing match by chance
r3	Arbitrary record, *r*	as method *r1* but recording length of search
r4	Arbitrary record, *r*	searching database of known individuals and finding unique match
r5	Arbitrary record, *r*	extending method *r1* by searching for additional matches
r1u, r2u, r3u,…	Arbitrary sample unique record, *r*	as for method *r1, r2, r3…* respectively
Directed Searches		
B1	Known individual in population, *B*	searching microdata records and finding unique match

Table 3. Estimated Identification Risk for Sample Unique Cases with Minimum, Median and Maximum Values of Estimated Cell Probabilities.

Combination of key variables, x	$\hat{p}_x$	F_x	**Estimated identification risk for different search methods**			
			method *r1* expression (5)	**method *r1u*** expression (10)	**method *r2*** expression (7).	**method *B1*** expression (9)
Area B, woman, aged 20-24, separated, Bangladeshi ethnic group, looking after home	1.56×10^{-7} minimum	5	0.9298	0.9301	0.8715	0.8721
Area A, man, aged 65-69, divorced, white British ethnic group, full-time employed	7.08×10^{-5} median	65	0.0149	0.0150	0.0148	0.0148
Area A, man, aged 40-44, re-married, white British ethnic group, full-time employed	0.00121 maximum	870	0.0009	0.0009	0.0009	0.0009