A software package for the application of probabilistic anonymisation to sensitive individual-level data: a proof of principle with an example from the ALSPAC birth cohort study

Demetris Avraam[1,2], Andy Boyd[2*], Harvey Goldstein[3,4], Paul Burton[1,2]

[1] Institute of Health and Society, Newcastle University, Newcastle, UK

[2] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[3] Graduate School of Education, University of Bristol, Bristol, UK

[4] Institute of Child Health, University College London, London, UK

*Corresponding author: a.w.boyd@bristol.ac.uk

Abstract

Individual-level data require protection from unauthorised access to safeguard confidentiality and security of sensitive information. Risks of disclosure are evaluated through privacy risk assessments and are controlled or minimised before data sharing and integration. The evolution from 'Micro Data Laboratory' traditions (i.e. access in controlled physical locations) to 'Open Data' (i.e. sharing individual-level data) drives the development of efficient anonymisation methods and protection controls. Effective anonymisation techniques should increase the uncertainty surrounding re-identification while retaining data utility; allowing informative data analysis. 'Probabilistic anonymisation' is one such technique, which alters the data by addition of random noise. In this paper, we describe the implementation of one probabilistic anonymisation technique into an operational software written in R and we demonstrate its applicability through application to analysis of asthma-related data from the ALSPAC cohort study. The software is designed to be used by data managers and users without the requirement of advanced statistical knowledge.

Keywords

Probabilistic anonymisation, disclosure control, measurement error, *h*-rank index, ALSPAC

Introduction

Data custodians managing longitudinal study data resources use a variety of policies and processes to manage risks to participant confidentiality and data security when sharing data. This can form a means to help meet legal requirements and also a component of wider strategies to retain participant trust and the public acceptability of research (Carter, Laurie, & Dixon-Woods, 2015). Approaches range from: 1) removing directly identifiable information (see Panel 1 for term definitions); 2) only providing access to accredited users; 3) allocating (project specific) pseudo IDs to each subject; 4) making adjustments to outlying values and small cell counts; 5) sub-setting datasets to only include data required for specific investigations; 6) transforming data through complex statistical processes that mask or block access to the underlying individual-level data; and, 7) sharing and using data within secure policy and procedural frameworks (Elliot, Mackey, O'Hara, & Tudor, 2016), such as Data Safe Havens (Burton et al., 2015).

The EU General Data Protection Regulation (GDPR) (European Parliament, 2018), through national implementations such as the Data Protection Act 2018 (DPA) (UK Parliament, 2018), distinguishes between personal data and anonymous data. Personal data is defined as "information relating to natural persons who: a) can be identified or who are identifiable, directly from the information in question; or b) who can be indirectly identified from that information in combination with other information". Therefore, personal information, includes data with direct identifier variables or data where identity can be determined through linking to other readily available information. This classification is important as the safeguards required for the use of personal information are far more stringent than the safeguards required for the use of anonymous data. The DPA - even when research exemptions apply - requires that individuals are informed of the use of their personal information, and that the security of the data is maintained through the research process. Furthermore, even when these safeguards are in place, the DPA requires that data are de-identified as soon in the research process as possible – ideally prior to the point when the data are provided to researchers. In contrast, anonymous data do not fall under the scope of the DPA (or GDPR) and are therefore exempt from these requirements.

Panel 1: Disclosure Control Terminology

Data Custodians: authorised individuals/entities who manage and share study data. While (typically) authorised to view identifiable data, there is a risk they can accidentally disclose data through data breaches or accidentally and spontaneously identify a participant.

Accredited User: a bona-fide professional working for a bona-fide institution for a bona-fide purpose who can be expected to operate professionally and to not deliberately disclose information. The potential for accidental disclosure remains. Similar to the 'Safe Researcher' concept.

External Attacker: an individual who will attempt to deliberately disclose participant information for malicious means.

Individual of Interest: the participant(s) targeted by an external attacker.

Direct Identifier(s): a data item which on its own could identify an individual (e.g. name, full date of birth, full address, health or other service ID number). The GDPR/DPA 2018 has expanded the legal definition of personal identifiers to include genetic sequence information (when used for linkage) as well as digital network identifiers (such as internet 'IP' addresses).

Indirect-Identifier(s) (aka Quasi-Identifiers): social or health variables with (context specific) potential to disclose an individual's identity (i.e. they are likely to be known or discoverable to an external attacker or spontaneously recognisable to someone who knows the individual), for example: parity, height, weight, disease status, occupation categories.

Non-Identifier(s): variables with exceptionally limited potential to disclose an individual's identity. These will tend to be transient values (e.g. blood pressure readings).

Under the new DPA 2018 legislation, longitudinal research studies are required (Article 35 of the GDPR) to consider the risks associated with data processing and use. Through conducting 'Data Protection Impact Assessments', data custodians will assess risks (e.g. loss of control of data when sharing with external research users) and will have to implement controls to mitigate these risks (e.g. effectively anonymising the data). Given the pressures to share data, it seems inevitable that DPA 2018 will provide a new impetus for data guardians to explore options for effective disclosure control. As a community, the data guardians of longitudinal studies should work together to understand the options available, the impact these may have on research utility and how to implement anonymisation strategies effectively. The risk of not doing this, is that poorly executed anonymisation strategies reveal sensitive information about participants and bring the research community into disrepute. While we are fortunate that there are no known examples of this within the longitudinal research community, we

should take note of parallel examples of poor practice (e.g., in 2014 the New York Taxi & Limousine Commission released data on 173 million individual journey's, yet a poor anonymisation strategy meant that individuals could easily be re-identified and their sensitive information breached (Pandurangan, 2014)).

Achieving anonymity in a dataset is challenging and is complicated by the fact that much population discovery science, particularly that informed by longitudinal studies, relies on broad data sets of granular detailed individual-level data. These data are ideal for assessing life-course associations and controlling for socially mediated status, yet are also ideally suited – given their rich and typically unique patterns of values – for identifying participants' real-world identities. This situation is further complicated by the fact that some indirect identifiers have research value (e.g. age, gender), so that the different classes of identifiers (direct, indirect, non) often cannot be viewed in isolation and that identification risk is context specific. Existing approaches to controlling for this risk, such as *k*-anonymisation (El Emam & Dankar, 2008; Sweeney, 2002), attempt to mask these patterns of uniqueness through supressing and aggregating data values. While this technique offers some protection to disclosure risk (Domingo-Ferrer, Sebé, & Castellà-Roca, 2004), it also has the potential to impact the utility of the data to inform the research question.

Goldstein and Shlomo (2018) suggest the use of a probabilistic anonymisation approach to perturb the data through the addition of random noise to some or all variables in the dataset. In this approach, the risks posed by an external 'attacker' who wished to re-identify an individual of interest from a dataset, are assessed. In this risk scenario, it is assumed that the attacker independently knows the individual's data values for some or all the identifying variables within the dataset. Using this information, the attacker could 'link' to the target individual's record using the unique patterns in their data, and therefore learn new information about that person from their associated attribute variables. To avoid such identification Goldstein and Shlomo propose that sufficient noise is generated and added to the identifying variables to disguise their values as they appear to any attacker. From the research perspective, the accredited user is provided with sufficient information to remove the effects of the noise during the analysis stage to recover the underlying data structure and

therefore to produce consistent parameter estimates. This is done through the use of statistical techniques for fitting models with measurement errors (see (Goldstein, Browne, & Charlton, 2017)).

This paper, in contrast to Goldstein and Shlomo's methodological manuscript, presents a pragmatic perspective with worked examples. To apply Goldstein and Shlomo's methodology, we have written an operational software package using the open-source statistical programming language R. We use data from participants of the Avon Longitudinal Study of Parents and Children (ALSPAC) birth cohort study to demonstrate the feasibility and practicality of the approach. For illustration of the method, we anonymise asthma-related data by adding differing degrees of noise. We then perform three exemplar analyses on the differing versions of anonymised data, treating the noise as measurement error. Finally, we assess how well the true model parameters are retrieved and we compare the differing risks of residual disclosure in the different data sets.

Software package

We have developed two functions in R (Avraam, 2018); the function *probAnon()* which adds noise to an input dataset, and the function *hRanks()* which generates a re-identification risk measure. Software to carry out the data modelling has been written in MATLAB (Mathworks, 2016).

The probAnon() function

The function *probAnon()*, applies probabilistic anonymisation to an input dataset. The algorithm first separates the input data into two subset data frames, one for the continuous (numerical) and one for the categorical (integer or factor) variables. Then, normally distributed random noise with user-specified variances is added, independently, to the continuous and categorical variables. For continuous variables, the variance of noise is specified as a percentage of each variable's observed variance in the argument *weights*, which is a vector ($w_1, \ldots, w_s$) of length s, where s is the number of continuous variables in the input

dataset. If the user does not specify the vector of weights, each weight is set to 0.1 by default which means that the variance of the added noise is equal to the 10% of the observed variance of the variable. The random noise added to each binary variable follows a normal distribution with zero mean and variance specified by the user. The added noise therefore, converts binary data to continuous. For the 'noisy' continuous form of 0-1 binary variables, the algorithm then truncates any negative values to 0 and any values greater than 1 to 1. This step is not strictly necessary particularly since it tends to increase identifiability risk but may be convenient for presentational purposes and serves in the present context to present a 'worse case' scenario. The output is then the input dataset plus the added noise. In addition, the argument *seed*, allows the user to set a certain random number generator. If this argument is not specified, the function bases the *seed* parameter on the local time (as determined by the computer's internal clock).

The hRanks() function

The function *hRanks()* calculates a re-identification risk measure (the *h*-rank index) of anonymised data using the method proposed by Goldstein and Shlomo (2018). The function takes as input arguments the original dataset and the anonymised dataset (both having the same dimensions). The conceptual basis of the *h*-rank index is to estimate the probability of an attacker being successful in identifying their individual of interest within the anonymised dataset. We assume that a potential attacker will have access to some information about an individual they are targeting (we note that this assumption is also explicit within Data Protection legislation and represents a data guardian's 'worst case' scenario).

We describe the logic of this function here and illustrate this in Panel 2. Initially (step 1), the algorithm calculates the Euclidean distances (defined as the square root of the sum of the squares of the differences between the corresponding coordinates of two vectors) between each row in the true dataset and all rows in the noisy dataset (i.e. a pair-wise comparison that ultimately assesses all possible pairs). It then (step 2), ranks the distances to determine how close each true record is to every record in the noisy dataset (i.e. a 1 to *n* comparison where *n* is the total number of records), and identifies the position of the closest record (i.e. the

record which corresponds to rank equal to one). We use the standard competition ranking method (where ties are allocated the same rank, and the next allocated rank is offset by the number of ties, e.g. "1224"), which is performed by the R function *rank()* with the argument *ties.method='min'*. In step 3, the algorithm generates a duplicate copy of the true dataset and computes the Euclidean distances between each row in the true dataset with all rows in the copy of the true dataset and ranks them in order of distance (i.e. a 1 to *n* comparison similar to step 2). In step 4, the algorithm identifies the ranks of the distances calculated in step 3 at the locations specified in step 2. Finally (step 5), the algorithm calculates the difference (*h*-rank index) between the ranks located in step 4 and the ranks located in step 2, and returns a vector with those differences. Note that the critical observations to identify in step 2 are all ranked 1 (or a tied equivalent) as the role of step 2 is to search for the closest noisy record to each true record. If *h*=0 for any one record that an attacker has available (and belongs to the dataset), then this implies that the noisy record identified by the attacker as the closest one in terms of the distance metric, is in fact the true one. The average value of *h*-rank indices provides a metric of disclosiveness. The larger the average value of *h*, the greater the level of unreliability in any attempt to disclose identity through exploiting a given individual's known pattern of data values. Where the average of *h* is small (i.e. lower than an acceptable threshold pre-specified by the study data custodian), the *probAnon()* function can be re-used to alter the data with a higher level of noise in order to increase the uncertainty of re-identification.

Some care is needed where there are more than a negligible number of tied distances. This will be a particular issue with categorical, including binary, data. For example, where a dataset consists of only four binary indicators, there are only 16 possible patterns; meaning that for any given record where noise has been added there will be many tied rank distances. For an attacker, when estimating *h*, this will result in additional uncertainty. Thus, for example, if there are *p* tied ranks and the correct true record is among these, the attacker will be confronted with *p* records with *h*=0, and will be able to choose the correct one only with probability 1/*p*. To reflect this so that we can consistently report our risk measure on the scale of *h*, a very small amount of noise is added to each of the identifiers in order to break the ties and so that the true record will therefore be identified as the closest with probability 1/*p.* In

the present implementation (for the case that we have a dataset with only categorical variables) we have added, for all categorical (binary) variables, noise following a normal distribution with variance 10^{-8}.

Panel 2: Schematic illustration of *probAnon()* and *hRanks()* algorithms

In this schematic illustration we use a hypothetical dataset containing three variables (var1 – var3) – two are continuous and one binary - relating to six data subjects (1-6). Our aim is to use the *probAnon()* function to generate 'noisy' version of the true records, and then to use *hRanks()* function to determine the strength of the disclosure control through quantifying the similarity of the noisy data to the true data via h.

probAnon()

In this example, the *probAnon()* function adds normal noise with zero mean and variance equal to 0.1% of the true variability to each continuous variable and normal noise with zero mean and variance equal to 0.5 to the binary variable. The binary variable after the addition of noise is returned as continuous where negative values are truncated to 0 and values greater than 1 are truncated to 1.

TRUE RECORDS

	var1	var2	var3
1	9.63	3.84	0
2	10.39	3.69	0
3	9.76	3.95	1
4	9.77	4.21	0
5	9.5	3.61	0
6	9.93	3.35	1

NOISY RECORDS

	var1	var2	var3
1	9.58	3.88	0.28
2	10.37	3.57	0.08
3	9.91	3.89	0.61
4	9.78	4.17	1
5	9.51	3.72	0.35
6	10.1	3.38	0

hRanks()

Step 1: The Euclidean distances between each row in the true dataset and each row in the noisy dataset are calculated (i.e. a pair-wise comparison that ultimately assesses all possible pairs). It then ranks the distances true-record-wise (in Step 2) in order to determine how close each individual's true record is to every noisy record. It is assumed that the potential attacker has access to one or more recorded true values.

Distances

TRUE RECORD \ NOISY RECORD	1	2	3	4	5	6
1	0.2872	0.7918	0.6731	1.0637	0.3890	0.6576
2	0.8778	0.1456	0.8016	1.2659	0.9475	0.4245
3	0.7455	1.1674	0.4221	0.2209	0.7334	1.2002
4	0.4727	0.8809	0.7029	1.0008	0.6559	0.8932
5	0.3971	0.8746	0.7865	1.1798	0.3670	0.6426
6	0.9601	1.0433	0.6664	0.8336	0.8578	1.0148

Step 2: Ranks (lowest to highest) of distances row-wise. The position of the lowest distance (circled number) indicates which of the noisy records is the closest to each true record.

→	①	5	4	6	2	3
→	4	①	3	6	5	2
→	4	5	2	①	3	6
→	①	4	3	6	2	5
→	2	5	4	6	①	3
→	4	6	①	2	3	5

Step 3: The Euclidean distances between each row in the true dataset (any one of which may be in the possession of the attacker) and each row in its duplicated copy dataset are calculated (i.e. a pair-wise comparison for all possible pairs). It then ranks the distances row-wise (Step 4).

TRUE RECORD \ TRUE RECORD	1	2	3	4	5	6
1	0	0.7747	1.0144	0.3956	0.2642	1.1533
2	0.7747	0	1.2102	0.8092	0.8936	1.1520
3	1.0144	1.2102	0	1.0333	1.0877	0.6236
4	0.3956	0.8092	1.0333	0	0.6580	1.3286
5	0.2642	0.8936	1.0877	0.6580	0	1.1192
6	1.1533	1.1520	0.6236	1.3286	1.1192	0

Step 4: Enclosed numbers are set to the positions identified in Step 2.

→	①	4	5	3	2	6
→	2	①	6	3	4	5
→	3	6	1	④	5	2
→	②	4	5	1	3	6
→	2	4	5	3	①	6
→	5	4	②	6	3	1

Step 5: The h-rank index is calculated as the difference of each circled rank of each individual record, selected in Step 4 and its corresponding (i.e. having the same position) circled rank located in Step 2. The vector of the h-rank indices for these six individual-level records is then $h = (0,0,3,1,0,1)$.

By carrying out the random noise addition many times, we obtain a distribution of h, so that we can estimate the cumulative distributions as given in the examples. We can further, evaluate these distributions for records at different percentiles of the distribution of distances from the centroid of the joint distribution of the identifiers.

Examples

To demonstrate the feasibility of the software we show its applicability to childhood asthma data from participants in ALSPAC; a longitudinal birth cohort study collecting information of participants' life-course exposures, and health, social and well-being outcomes (Boyd et al., 2013). ALSPAC recruited pregnant women living in, and around, the City of Bristol (south west UK) – who were due to deliver between 01/04/91 and 31/12/92. An initial total of 14,062 live-born children were enrolled. By age 18, the enrolled sample had extended to include 14,775 live-born individuals from 15,247 pregnancies. The assessment in this paper was conducted on a sample of 15,211 participants. Data is collected via questionnaires, study assessment visits, biological and 'omic characterisations and linkage to routine records (see: 'www.bristol.ac.uk/alspac/researchers/access/' for ALSPAC data dictionary).

Ethical approval for ALSPAC was obtained from the ALSPAC Law & Ethics Committee and the NHS Research Ethics Committees. The variables used in this exemplar application (see Table 1) were selected and then reviewed by an ALSPAC data custodian (author AB) using the ALSPAC privacy impact / risk assessment template. This assessment (based on an assumption that a potential attacker had some access to real information about their target) noted that the dataset contained Direct Identifiers (study ID), Indirect Identifiers, Non-Identifiers and Outcome variables (see Table 1).

Table 1: Asthma-related variables from the ALSPAC birth cohort study.

Variable identification	Type	Identifier/ Outcome	Missing values*	Explanation
b650	binary	Indirect	2009	ever smoked (completed by mother at 18 weeks of gestation)
kz021	binary	Indirect	517	child's sex
kc362	binary	Indirect	4144	never exposed to passive smoke (completed by mother at 15 months)
kc401	multi-categorical	Indirect	4231	ever breast fed (completed by mother at 15 months)
m2110	binary	Non	7036	there is damp/condensation/ mould in home (completed by mother at 7 years 1 month)

dda_91	binary	Outcome	7053	doctor ever diagnosed asthma (completed by mother at 91 months)
kv1059	multi-categorical	Outcome	7426	child had asthma in past 12 months (completed by mother at 128 months)
height_f8	continuous	Indirect	8028	child's height (cm), (measured by fieldworker at 'focus@8' clinical assessment visit at mean age 103.8 months)
weight_f8	continuous	Indirect	8249	child's weight (kg), (measured by fieldworker at 'focus@8' clinical assessment visit at mean age 103.8 months)
raw_fev1_f8	continuous	Outcome	8301	forced expiratory volume in 1 second, (measured by fieldworker at 'focus@8' clinical assessment visit at mean age 103.8 months)

*The number of missing values includes also the 'not completed', 'don't know' and 'no response' answers.

We conducted a complete case analysis that was restricted to participants with non-missing data on all relevant variables. We calculated the children's Body Mass Index (BMI) using the relationship $\mathrm{BMI} = \mathrm{weight} \,/\, (\mathrm{height}/100)^2$. We then created three separate datasets: dataset A with the variables $\mathrm{ASTHMA} = \mathrm{dda\_91}$, $\mathrm{SMOKE} = \mathrm{kc362}$, $\mathrm{BREAST\ FED} = \mathrm{kc401}$ and $\mathrm{MOULD} = \mathrm{m2110}$; dataset B with the variables $\mathrm{ASTHMA} = \mathrm{dda\_91}$, $\mathrm{BMI} = \mathrm{weight\_f8}/(\mathrm{height\_f8})^2$ and $\mathrm{BREAST\ FED} = \mathrm{kc401}$; dataset C with the variables $\mathrm{FEV1} = \mathrm{raw\_fev1\_f8}$, $\mathrm{BMI} = \mathrm{weight\_f8} \,/\, (\mathrm{height\_f8})^2$, $\mathrm{SEX} = \mathrm{kz021}$ and $\mathrm{SMOKE} = \mathrm{b650}$. From each of the three datasets we removed any rows with missing values. This results in datasets with 6837, 4975 and 5942 complete records respectively. We then converted the multi-categorical variables (see type of each variable in Table 1) to binary data. For the variable kc401 (ever breast fed), we combined together the categories "Yes, no longer" and "Yes, still" and replaced their values with ones while we replaced the values in the category "No, never" with zeros. For the variable kv1059 (child had asthma in past 12 months, completed by mother at 128 months), we combined together the categories "Yes, but did not see a doctor" and "Yes, saw a doctor" and replaced their values with ones while we replaced the values in the category "No, did not have" with zeros. We finally generated 'noisy' datasets for each true dataset (A-C) using the *probAnon()* function. For datasets A and B we do not add noise to the ASTHMA variable which

is used as the response variable in a probit regression model. For all the other variables we add normally distributed noise with zero mean and variance equal to the value shown in Table 2. We have not considered the case where noise is added to the response variable where this is binary. This feature is described in Goldstein and Shlomo (2018) but is not yet been implemented in the analysis software.

Table 2: Variances of the added noise. Note that for binary variables we add different levels of noise.

Dataset	Variable	Type	True variance of variable	Variance of added noise
A	ASTHMA	binary response	0.160	-
	SMOKE	binary covariate	0.231	0.05, 0.1, 0.2, 0.5
	BREAST FED	binary covariate	0.171	0.05, 0.1, 0.2, 0.5
	MOULD	binary covariate	0.244	0.05, 0.1, 0.2, 0.5
B	ASTHMA	binary response	0.109	-
	BMI	continuous covariate	5.512	0.55
	BREAST FED	binary covariate	0.151	0.05, 0.1, 0.2, 0.5
C	FEV1	continuous response	0.069	0.0069
	BMI	continuous covariate	5.828	0.58
	SEX	binary covariate	0.250	0.05, 0.1, 0.2, 0.5
	SMOKE	binary covariate	0.248	0.05, 0.1, 0.2, 0.5

Results

We apply regression models to the true and noisy data of each dataset (A-C) and compare the estimated coefficients. Each regression model is applied to the true data using common functions for generalised linear models (e.g. *glm()* function in R) and to the noisy data using a Bayesian Markov Chain Monte Carlo (MCMC) algorithm that allows the recovery of the original data structure (see description of this procedure in (Goldstein et al., 2017)). We have not run full simulations of the data. We note that a true simulation to derive population estimates will require both the generation of a model using assumed population parameters and for each of these generated datasets the further generation of a set of models where the noise is sampled from the assumed noise distribution. Goldstein and Shlomo (2018) ran

simulations with both continuous and binary covariates that pointed to negligible bias for the general procedure.

Dataset A

Dataset A consists only of binary data and without special care will give many tied rank distances. To illustrate this, we present the frequencies of all possible combinations for the true values of Dataset A in table 3.

Table 3. Frequencies for all possible combinations of values for dataset A.

ASTHMA	SMOKE	BREAST FED	MOULD	Frequency
0	0	0	0	380
0	0	0	1	181
0	0	1	0	1691
0	0	1	1	1308
0	1	0	0	357
0	1	0	1	227
0	1	1	0	758
0	1	1	1	567
1	0	0	0	91
1	0	0	1	61
1	0	1	0	348
1	0	1	1	310
1	1	0	0	133

We see from table 3 that the smallest set of identical combinations of identifiers confronting an attacker is 61, and the largest 1691. We have therefore used the procedure of adding additional 'tie-breaking' noise and see that the probabilities for a successful attack are still acceptably small. Table 4 gives the estimates of disclosiveness as expressed in terms of h based on 100 simulations. The big number of simulations is used to demonstrate stable estimates.

Table 4: Cumulative probabilities of *h* for noisy dataset A, at 10th, 50th and 90th percentiles. No noise is added to the response variable. Noise with variance 0.05 (scenario 1), 0.1 (scenario 2), 0.2 (scenario 3) and 0.5 (scenario 4) is added to all predictors. 100 simulated noise additions used.

Scenario	Percentile	$\mathrm{P}(h=0)$	$\mathrm{P}(h\leq 1)$	$\mathrm{P}(h\leq 2)$	$\mathrm{P}(h\leq 3)$	$\mathrm{P}(h\leq 4)$	$\mathrm{P}(h\leq 5)$
	10%	0.0265	0.0286	0.0307	0.0327	0.0343	0.0369
1	50%	0.0105	0.0128	0.0155	0.0177	0.0200	0.0219
	90%	0.0088	0.0112	0.0139	0.0162	0.0185	0.0208
	10%	0.0230	0.0250	0.0275	0.0301	0.0326	0.0340
2	50%	0.0091	0.0116	0.0141	0.0167	0.0191	0.0213
	90%	0.0069	0.0094	0.0119	0.0143	0.0167	0.0191
	10%	0.0221	0.0239	0.0253	0.0273	0.0281	0.0300
3	50%	0.0084	0.0109	0.0129	0.0150	0.0169	0.0193
	90%	0.0063	0.0087	0.0110	0.0132	0.0151	0.0174
	10%	0.0143	0.0156	0.0165	0.0175	0.0182	0.0189
4	50%	0.0048	0.0063	0.0076	0.0088	0.0102	0.0113
	90%	0.0034	0.0049	0.0063	0.0078	0.0092	0.0104

To analyse the data from Dataset A, we apply a probit regression model where the asthma indicator is regressed on smoking, breast feeding and presence of mould

$$\mathrm{probit}(ASTHMA) = \beta_0 + \beta_1(SMOKE) + \beta_2(BREAST\ FED) + \beta_3(MOULD). \quad (1)$$

The estimated coefficients from the analysis, and their standard errors, are shown in Table 5. We observe that the estimates of the model applied to noisy data using the procedure that removes the noise are close to the estimates of the model applied to the true data (i.e. an overlap between their confidence intervals).

Table 5: Estimated parameters and their standard errors for dataset A. The first row shows the estimated coefficients of the model applied to the true records and scenarios 1-4 show the estimated coefficients of the model applied to noisy data using procedures to recover the data structure. Note that the response variable was without noise and noise with variance 0.05 (scenario 1), 0.1 (scenario 2), 0.2 (scenario 3) and 0.5 (scenario 4) was added to all predictors. The results in scenarios 1-4 show the means of 50 MCMC simulations.

Scenario	Data	β_0 (SE)	β_1 (SE)	β_2 (SE)	β_3 (SE)
	True data	-0.809 (0.043)	0.124 (0.036)	-0.131 (0.042)	0.053 (0.035)
1	Noisy data	-0.817 (0.018)	0.115 (0.032)	-0.114 (0.038)	0.048 (0.031)
2	Noisy data	-0.777 (0.042)	0.115 (0.043)	-0.167 (0.044)	0.048 (0.036)
3	Noisy data	-0.828 (0.031)	0.062 (0.027)	-0.063 (0.026)	0.027 (0.026)

4	Noisy data	-0.842 (0.024)	0.040 (0.019)	-0.027 (0.020)	0.016 (0.019)

Dataset B

Dataset B includes both binary and continuous covariates. We add noise with variance equal to 10% of the true variance to BMI variable and noise with variance 0.05, 0.1, 0.2 and 0.5 to the breast feeding variable. The cumulative probabilities of *h* based on 100 simulations are shown in Table 6. We observe that probabilities of *h* to be less than a certain value are increasing with the increase of noise added to the binary variable (i.e. comparing the values between scenario 1 which refers to noise with variance 0.05 and scenario 4 which refers to noise with variance 0.5 added to the binary breast feeding variable). We also observe higher values of probabilities at the 10th and 90th percentiles in contrast with the lower values at the median (50th percentile). In addition, the probabilities at the 10th percentile are systematically higher than the probabilities at the 90th percentile which is related to the slightly right-skewed actual distribution of the continuous BMI.

Table 6: Cumulative probabilities of *h* for noisy dataset B, at 10th, 50th and 90th percentiles. No noise is added to the response variable. Noise with variance 0.55 was added to BMI and noise with variance 0.05 (scenario 1), 0.1 (scenario 2), 0.2 (scenario 3) and 0.5 (scenario 4) was added to breast feeding variable. 100 simulated noise additions used in computations.

Scenario	Percentile	$\mathrm{P}(h=0)$	$\mathrm{P}(h\leq 1)$	$\mathrm{P}(h\leq 2)$	$\mathrm{P}(h\leq 3)$	$\mathrm{P}(h\leq 4)$	$\mathrm{P}(h\leq 5)$
	10%	0.0071	0.0136	0.0191	0.0246	0.0299	0.0356
1	50%	0.0060	0.0115	0.0168	0.0219	0.0269	0.0319
	90%	0.0070	0.0128	0.0183	0.0234	0.0285	0.0336
	10%	0.0071	0.0136	0.0191	0.0247	0.0300	0.0358
2	50%	0.0059	0.0113	0.0167	0.0217	0.0267	0.0316
	90%	0.0069	0.0126	0.0181	0.0232	0.0283	0.0333
	10%	0.0066	0.0127	0.0178	0.0232	0.0282	0.0335
3	50%	0.0055	0.0106	0.0156	0.0203	0.0249	0.0296
	90%	0.0065	0.0118	0.0170	0.0218	0.0266	0.0314
	10%	0.0056	0.0106	0.0149	0.0196	0.0237	0.0283
4	50%	0.0047	0.0089	0.0131	0.0171	0.0210	0.0250
	90%	0.0055	0.0101	0.0145	0.0186	0.0227	0.0268

For dataset B, we apply a probit regression model where the asthma indicator is regressed on BMI and breast feeding

$$\text{probit}(ASTHMA) = \beta_0 + \beta_1(BMI) + \beta_2(BREAST\ FED). \quad (2)$$

A comparison of the results derived from the model applied to the true and the noisy data are shown in Table 7. Similarly to the results obtained from Dataset A, we observe a highly accurate estimation of the model parameters by fitting the model to the noisy data and removing the noise using MCMC procedures.

Table 7: Estimated parameters and their standard errors for dataset B. The first row shows the estimated coefficients of the model applied to the true records and scenarios 1-4 show the estimated coefficients of the model applied to noisy data where noise with variance 0.55 was added to BMI and noise with variance 0.05 (scenario 1), 0.1 (scenario 2), 0.2 (scenario 3) and 0.5 (scenario 4) was added to the breast feeding variable. Note that the response variable is without noise. The results in scenarios 1-4 show the means of 50 MCMC simulations.

Scenario	Data	β_0 (SE)	β_1 (SE)	β_2 (SE)
	True data	-1.261 (0.174)	0.010 (0.010)	-0.078 (0.058)
1	Noisy data	-1.256 (0.174)	0.009 (0.010)	-0.072 (0.054)
2	Noisy data	-1.250 (0.170)	0.008 (0.011)	-0.065 (0.070)
3	Noisy data	-1.325 (0.176)	0.011 (0.010)	-0.027 (0.036)
4	Noisy data	-1.330 (0.167)	0.011 (0.010)	-0.014 (0.026)

Dataset C

Probabilistic anonymisation has been also applied to Dataset C and the cumulative probabilities for *h* at different percentiles based on 100 simulations are shown in Table 8. The difference in Dataset C (in contrast with datasets A-B) is the outcome variable which is continuous instead of binary and therefore noise is added to the outcome in the same way as the noise is added to any continuous explanatory variables.

Table 8: Cumulative probabilities of *h* for noisy dataset C, at 10th, 50th and 90th percentiles. Noise with variance 0.0069 was added to the outcome FEV1 variable, noise with variance 0.58 was added to BMI and noise with variance 0.05 (scenario 1), 0.1 (scenario 2), 0.2 (scenario 3) and 0.5 (scenario 4) was added to sex and smoke indicators. 100 simulations used.

Scenario	Percentile	$\mathrm{P}(h=0)$	$\mathrm{P}(h\leq 1)$	$\mathrm{P}(h\leq 2)$	$\mathrm{P}(h\leq 3)$	$\mathrm{P}(h\leq 4)$	$\mathrm{P}(h\leq 5)$
1	10%	0.0204	0.0377	0.0522	0.0661	0.0785	0.0896
	50%	0.0206	0.0371	0.0514	0.0644	0.0761	0.0866

	90%	0.0209	0.0375	0.0522	0.0650	0.0768	0.0874
	10%	0.0194	0.0365	0.0505	0.0648	0.0772	0.0878
2	50%	0.0196	0.0356	0.0496	0.0625	0.0741	0.0845
	90%	0.0200	0.0361	0.0505	0.0631	0.0748	0.0852
	10%	0.0179	0.0332	0.0463	0.0591	0.0708	0.0807
3	50%	0.0180	0.0327	0.0457	0.0575	0.0685	0.0782
	90%	0.0183	0.0332	0.0465	0.0582	0.0691	0.0789
	10%	0.0140	0.0264	0.0365	0.0465	0.0561	0.0641
4	50%	0.0143	0.0258	0.0359	0.0453	0.0540	0.0615
	90%	0.0146	0.0264	0.0369	0.0461	0.0548	0.0625

For dataset C, the force expiratory volume in 1 second is regressed on BMI, sex and smoking

$$FEV1 = \beta_0 + \beta_1(BMI) + \beta_2(SEX) + \beta_3(SMOKE). \quad (3)$$

The estimated coefficients with their standard errors are shown in Table 9.

Table 9: Estimated parameters and their standard errors for dataset C. The first row shows the estimated coefficients of the model applied to the true records and scenarios 1-4 show the estimated coefficients estimated from the model applied to noisy data. Noise with variance 0.0069 was added to the outcome FEV1 variable, noise with variance 0.58 was added to BMI and noise with variance 0.05 (scenario 1), 0.1 (scenario 2), 0.2 (scenario 3) and 0.5 (scenario 4) was added to sex and smoking indicators. The results in scenarios 1-4 show the means of 50 MCMC simulations.

Scenario	Data	β_0 (SE)	β_1 (SE)	β_2 (SE)	β_3 (SE)
	True data	1.139 (0.024)	0.029 (0.001)	0.105 (0.007)	-0.005 (0.007)
1	Noisy data	1.145 (0.025)	0.029 (0.001)	0.096 (0.006)	-0.004 (0.006)
2	Noisy data	1.140 (0.026)	0.029 (0.001)	0.051 (0.032)	-0.005 (0.006)
3	Noisy data	1.182 (0.025)	0.028 (0.001)	0.051 (0.005)	-0.004 (0.005)
4	Noisy data	1.194 (0.025)	0.028 (0.001)	0.031 (0.004)	-0.001 (0.004)

We see from these example analyses that the disclosure risk increases with the number of identifying variables used, but remains acceptable. These results suggest that for similar datasets the amount of noise added could safely be reduced. Nevertheless, when a large number of variables is involved in a dataset, the values of *h*-rank index will be expected to increase and this is clearly an area for further exploration. We also note that, especially for binary variables, estimates derived from the noisy data can have large standard errors and

the true estimates from the real data can be very different. The amount of noise added to the binary variables in scenario 4 has a standard deviation $\sqrt{0.5} = 0.71$ which is very large compared to the range (0,1) of the true data and therefore we get a lot of instability (as 25% of observed zeros get wrongly defined to their true category). This suggests that further work is needed for such cases and Table 8 suggests that smaller values should produce acceptable values for *h*. We conclude that for even low levels of noise the method is sufficient to effectively anonymise the records, but we show the example of noise added to the binary with variance 0.5 as a warning to data managers on the increase in the loss of utility.

Discussion

We have shown how a probabilistic anonymisation procedure can be applied to data management procedures in such a way that disclosure risk is reduced to acceptable levels while retaining the ability to carry out statistical analysis. The analysis conducted on the noisy, anonymous, data suffered some loss of statistical efficiency when compared with analysis on the true data; a consequence of which is enlarged confidence intervals and fewer significant inferences. Where the variance of noise added to the binary covariates is large (i.e. > 0.2), there is likely to be unacceptably high loss of statistical efficiency and for binary data biases may also be introduced. This example illustrates the challenge in balancing disclosure control with retaining data utility.

When considering disclosure risk, data custodians should consider the risk of motivated external attackers, accidental disclosure to authorised users and also the possible consequences of human error. In the first scenario (external attacker) the attacker may be motivated to identify a given individual due to their notoriety (for example an investigative journalist following a story – or a researcher illustrating the fallacy of supposed 'anonymity' (Sweeney, 2002)) or out of personal interest. In the second, an accredited user or data custodian may recognise an individual during their legitimate work, and in the third an authorised user may inadvertently release a dataset to a wider than authorised audience through a data breach. In all these scenarios identification of a given data subject is likely to occur through matching known 'real world' information about an individual to equivalent

information about the same individual within a dataset. Probabilistic anonymisation helps control for these risks by removing certainty about whether the values being considered in the noisy data are true 'real world' values. The *h*-rank index disclosure measure proposed by Goldstein and Shlomo (2018), adopts this perspective by seeking to establish how well any single individual is 'hidden amongst the crowd' of the other individuals in the dataset. While this approach seems conceptually appropriate – it has some limitations. We found that, in its current state of development, the *h*-rank index was unable to adequately account for disclosure risk of outliers (this was acknowledged by Goldstein and Shlomo (2018) but this can be addressed through suitable pre-processing techniques such as truncating them. It was also unable to account for the disclosure risk of clusters of individuals who all have the same outcome value, i.e. that it is not necessary to identify the Individual of Interest within the cluster if they all have the same outcome of interest. This phenomenon is known elsewhere in the privacy literature (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006) and could be quantified by including *l*-diversity metrics to assess outcome value diversity. Finally we found that the *h*-rank index was also unable to account for the protective benefits of the sample being selected from a wider population (again a point noted by Goldstein and Shlomo (2018)). While this last point would be difficult to accommodate in a metric, it could be incorporated into the Data Custodians risk assessment process.

In practice, a data custodian, in conjunction with potential accredited users would need to evaluate the risks associated with applying any given amount of noise related to the potential loss of analysis efficiency. In some cases, it may not be desirable to release data into the public domain. We suggest that there are few, if any, situations where some variables of interest would need to be excluded; though this remains an area for further study. However, our findings that large amounts of noise impact model estimates suggest this may limit the application for data sets treated with large amounts of noise, e.g. they may be suitable only for training or data exploration rather than applied research. A more realistic application would be the use of probabilistic anonymisation to applying limited amounts of noise (to protect against spontaneous recognition or contained (i.e. not public) data breaches) and to supply accredited users with these noisy data under controlled 'safe haven' conditions. As such, probabilistic anonymisation will add to the range of tools available to data managers

that include manual data transformations (such as outlier suppression), statistical approaches (e.g. synthetic data and *k*-anonymisation) and distributed 'black box' computing approaches (e.g. DataSHIELD (Wilson et al., 2017)). All such approaches involve trade-offs between disclosure control, impact on utility and impact on usability. Probabilistic anonymisation has one clear advantage over some of these approaches (e.g. *k*-anonymisation and synthetic data) in that it allows efficient and accurate data linkage to additional datasets, given that the noisy data can contain ID numbers and the noise can be applied over multiple datasets independently. Further work is needed to assess the extent to which these trade-offs apply in order to help inform the Data Custodian community as to which approach may best fit any given situation.

The software functions developed here are proof of principle rather than fully developed 'commercial grade' software. We have identified that improvements would be needed in the following areas before wider adoption: 1) the code needs to accommodate multi-category categorical variables; 2) missing values are not currently supported, we need to allow for these or to develop alternative approaches (e.g. imputation); 3) the *h*-rank index needs developing (as described above) and further consideration given to accommodating outlying values in a flexible manner.

We have demonstrated that probabilistic anonymisation can be effectively deployed to help control for disclosure risk while producing accurate estimates We have assumed that the data to be used is for bona-fide research scenarios (i.e. not for releasing data into the public domain) where responsible and verifiable data security measures are in place. Additionally, this concept would be novel to many data custodians who may not have advanced statistical expertise, so that determining the appropriate balance between disclosure risk control and retaining data utility would require training. With the enhancements we have identified the software assessed here could be developed into a fully functional tool for Data Custodians. This software would be a useful tool to help longitudinal studies maintain participant trust and to share data securely and effectively while meeting ever more stringent data protection requirements.

Acknowledgements

We are extremely grateful to the families who took part in ALSPAC, the midwives who helped recruit them, and the whole ALSPAC team (including interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses). The UK Medical Research Council and the Wellcome Trust (102215/2/13/2) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors (DA developed the software and drafted the methods and results; AB drafted the introduction and discussion; all authors contributed to the final manuscript) who serve as guarantors. This research is part of a collaborative research programme entitled 'Cohorts and Longitudinal Studies Enhancement Resources' (CLOSER). This programme is funded by the UK Economic and Social Research Council and Medical Research Council ES/K000357/1. The funders took no role in the design, execution, analysis or interpretation of the data or in the writing up of the findings.

References

Avraam, D. (2018). R functionality for application of probabilistic anonymisation. Available in https://github.com/davraam/Probabilistic_Anonymisation/releases/tag/v1.0.0

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., . . . Davey Smith, G. (2013). Cohort Profile: the 'children of the 90s' - the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology, 42*(1), 111-127.

Burton, P. R., Murtagh, M. J., Boyd, A., Williams, J. B., Dove, E. S., Wallace, S. E., . . . Knoppers, B. M. (2015). Data Safe Havens in health research and healthcare. *Bioinformatics, 31*(20), 3241-3248.

Carter, P., Laurie, G. T., & Dixon-Woods, M. (2015). The social licence for research: why care.data ran into trouble. *Journal of Medical Ethics, 41*(5), 404-409.

Domingo-Ferrer, J., Sebé, F., & Castellà-Roca, J. (2004). On the Security of Noise Addition for Privacy in Statistical Databases. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in Statistical Databases: CASC Project Final Conference, PSD 2004, Barcelona, Spain, June 9-11, 2004. Proceedings* (pp. 149-161). Berlin, Heidelberg: Springer Berlin Heidelberg.

El Emam, K., & Dankar, F. K. (2008). Protecting Privacy Using k-Anonymity. *Journal of the American Medical Informatics Association, 15*(5), 627-637.

Elliot, M., Mackey, E., O'Hara, K., & Tudor, C. (2016). *The Anonymisation Decision-Making Framework*. Manchester, UK: UKAN.

European Parliament. (2018). General Data Protection Regulation. Retrieved from https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

Goldstein, H., Browne, W. J., & Charlton, C. (2017). A Bayesian model for measurement and misclassification errors alongside missing data, with an application to higher education participation in Australia. *Journal of Applied Statistics*, 1-14.

Goldstein, H., & Shlomo, N. (2018). *A Probabilistic Procedure for Anonymisation and Analysis of Perturbed Datasets*. Under Review. A copy can be obtained from Harvey Goldstein: h.goldstein@bristol.ac.uk

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). *L-diversity: privacy beyond k-anonymity.* Paper presented at the 22nd International Conference on Data Engineering (ICDE'06).

Mathworks. (2016). MATLAB. Retrieved from http://uk.mathworks.com/products/matlab/

Pandurangan, V. (2014). On Taxis and Rainbow Tables: Lessons for researchers and governments from NYC's improperly anonymized taxi logs. Retrieved from http://blogs.lse.ac.uk/impactofsocialsciences/2014/07/16/nyc-improperly-anonymized-taxi-logs-pandurangan/

Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), 557-570.

UK Parliament. (2018). Data Protection Act. Retrieved from http://www.legislation.gov.uk/ukpga/2018/12/pdfs/ukpga_20180012_en.pdf

Wilson, R. C., Butters, O. W., Avraam, D., Baker, J., Tedds, J. A., Turner, A., . . . Burton, P. R. (2017). DataSHIELD – New Directions and Dimensions. *Data Science Journal, 16*, 21.