nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-10933-3 OPEN

# Estimating the success of re-identifications in incomplete datasets using generative models

Luc Rocher [1,2,3], Julien M. Hendrickx[1] & Yves-Alexandre de Montjoye[2,3]

While rich medical, behavioral, and socio-demographic data are key to modern data-driven research, their collection and use raise legitimate privacy concerns. Anonymizing datasets through de-identification and sampling before sharing them has been the main tool used to address those concerns. We here propose a generative copula-based method that can accurately estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. On 210 populations, our method obtains AUC scores for predicting individual uniqueness ranging from 0.84 to 0.97, with low false-discovery rate. Using our model, we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model.

[1] Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium. [2] Department of Computing, Imperial College London, London SW7 2AZ, UK. [3] Data Science Institute, Imperial College London, London SW7 2AZ, UK. Correspondence and requests for materials should be addressed to Y.-A.d.M (email: deMontjoye@imperial.ac.uk)

In the last decade, the ability to collect and store personal data has exploded. With two thirds of the world population having access to the Internet[1], electronic medical records becoming the norm[2], and the rise of the Internet of Things, this is unlikely to stop anytime soon. Collected at scale from financial or medical services, when filling in online surveys or liking pages, this data has an incredible potential for good. It drives scientific advancements in medicine[3], social science[4,5], and AI[6] and promises to revolutionize the way businesses and governments function[7,8].

However, the large-scale collection and use of detailed individual-level data raise legitimate privacy concerns. The recent backlashes against the sharing of NHS [UK National Health Service] medical data with DeepMind[9] and the collection and subsequent sale of Facebook data to Cambridge Analytica[10] are the latest evidences that people are concerned about the confidentiality, privacy, and ethical use of their data. In a recent survey, >72% of U.S. citizens reported being worried about sharing personal information online[11]. In the wrong hands, sensitive data can be exploited for blackmailing, mass surveillance, social engineering, or identity theft.

De-identification, the process of anonymizing datasets before sharing them, has been the main paradigm used in research and elsewhere to share data while preserving people's privacy[12–14]. Data protection laws worldwide consider anonymous data as not personal data anymore[15,16] allowing it to be freely used, shared, and sold. Academic journals are, e.g., increasingly requiring authors to make anonymous data available to the research community[17]. While standards for anonymous data vary, modern data protection laws, such as the European General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), consider that each and every person in a dataset has to be protected for the dataset to be considered anonymous[18–20]. This new higher standard for anonymization is further made clear by the introduction in GDPR of pseudonymous data: data that does not contain obvious identifiers but might be re-identifiable and is therefore within the scope of the law[16,18].

Yet numerous supposedly anonymous datasets have recently been released and re-identified[15,21–31]. In 2016, journalists re-identified politicians in an anonymized browsing history dataset of 3 million German citizens, uncovering their medical information and their sexual preferences[23]. A few months before, the Australian Department of Health publicly released de-identified medical records for 10% of the population only for researchers to re-identify them 6 weeks later[24]. Before that, studies had shown that de-identified hospital discharge data could be re-identified using basic demographic attributes[25] and that diagnostic codes, year of birth, gender, and ethnicity could uniquely identify patients in genomic studies data[26]. Finally, researchers were able to uniquely identify individuals in anonymized taxi trajectories in NYC[27], bike sharing trips in London[28], subway data in Riga[29], and mobile phone and credit card datasets[30,31].

Statistical disclosure control researchers and some companies are disputing the validity of these re-identifications: as datasets are always incomplete, journalists and researchers can never be sure they have re-identified the right person even if they found a match[32–35]. They argue that this provides strong plausible deniability to participants and reduce the risks, making such de-identified datasets anonymous including according to GDPR[36–39]. De-identified datasets can be intrinsically incomplete, e.g., because the dataset only covers patients of one of the hospital networks in a country or because they have been subsampled as part of the de-identification process. For example, the U.S. Census Bureau releases only 1% of their decennial census and sampling fractions for international census range from 0.07% in India to 10% in South American countries[40]. Companies are adopting similar approaches with, e.g., the Netflix Prize dataset including <10% of their users[41].

Imagine a health insurance company who decides to run a contest to predict breast cancer and publishes a de-identified dataset of 1000 people, 1% of their 100,000 insureds in California, including people's birth date, gender, ZIP code, and breast cancer diagnosis. John Doe's employer downloads the dataset and finds one (and only one) record matching Doe's information: male living in Berkeley, CA (94720), born on January 2nd 1968, and diagnosed with breast cancer (self-disclosed by John Doe). This record also contains the details of his recent (failed) stage IV treatments. When contacted, the insurance company argues that matching does not equal re-identification: the record could belong to 1 of the 99,000 other people they insure or, if the employer does not know whether Doe is insured by this company or not, to anyone else of the 39.5M people living in California.

Our paper shows how the likelihood of a specific individual to have been correctly re-identified can be estimated with high accuracy even when the anonymized dataset is heavily incomplete. We propose a generative graphical model that can be accurately and efficiently trained on incomplete data. Using socio-demographic, survey, and health datasets, we show that our model exhibits a mean absolute error (MAE) of 0.018 on average in estimating population uniqueness[42] and an MAE of 0.041 in estimating population uniqueness when the model is trained on only a 1% population sample. Once trained, our model allows us to predict whether the re-identification of an individual is correct with an average false-discovery rate of <6.7% for a 95% threshold ($\widehat{\xi_x} > 0.95$) and an error rate 39% lower than the best achievable population-level estimator. With population uniqueness increasing fast with the number of attributes available, our results show that the likelihood of a re-identification to be correct, even in a heavily sampled dataset, can be accurately estimated, and is often high. Our results reject the claims that, first, re-identification is not a practical risk and, second, sampling or releasing partial datasets provide plausible deniability. Moving forward, they question whether current de-identification practices satisfy the anonymization standards of modern data protection laws such as GDPR and CCPA and emphasize the need to move, from a legal and regulatory perspective, beyond the de-identification release-and-forget model.

## Results

**Using Gaussian copulas to model uniqueness**. We consider a dataset $\mathcal{D}$, released by an organization, and containing a sample of $n_{\mathcal{D}}$ individuals extracted at random from a population of $n$ individuals, e.g., the US population. Each row $\boldsymbol{x}^{(i)}$ is an individual record, containing $d$ nominal or ordinal attributes (e.g., demographic variables, survey responses) taking values in a discrete sample space $\mathcal{X}$. We consider the rows $\boldsymbol{x}^{(i)}$ to be independent and identically distributed, drawn from the probability distribution $X$ with $\mathbb{P}(X = \boldsymbol{x})$, abbreviated $p(\boldsymbol{x})$.

Our model quantifies, for any individual $\boldsymbol{x}$, the likelihood $\xi_{\boldsymbol{x}}$ for this record to be unique in the complete population and therefore always successfully re-identified when matched. From $\xi_{\boldsymbol{x}}$, we derive the likelihood $\kappa_{\boldsymbol{x}}$ for $\boldsymbol{x}$ to be correctly re-identified when matched, which we call correctness. If Doe's record $\boldsymbol{x}^{(d)}$ is unique in $\mathcal{D}$, he will always be correctly re-identified ($\kappa_{\boldsymbol{x}^{(d)}} = 1$ and $\xi_{\boldsymbol{x}^{(d)}} = 1$). However, if two other people share the same attribute ($\boldsymbol{x}^{(d)}$ not unique, $\xi_{\boldsymbol{x}^{(d)}} = 0$), Doe would still have one chance out of three to have been successfully re-identified ($\kappa_{\boldsymbol{x}^{(d)}} = 1/3$). We model $\xi_{\boldsymbol{x}}$ as:

$$\xi_{\boldsymbol{x}} \equiv \mathbb{P}\left(\boldsymbol{x} \text{ unique in } (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}) \mid \exists i, \boldsymbol{x}^{(i)} = \boldsymbol{x}\right) \quad (1)$$

$$= (1 - p(\boldsymbol{x}))^{n-1} \quad (2)$$

and $\kappa_{\boldsymbol{x}}$ as:

$$\kappa_{\boldsymbol{x}} \equiv \mathbb{P}\Big(\boldsymbol{x} \text{ correctly matched in } (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}) \mid \exists i, \boldsymbol{x}^{(i)} = \boldsymbol{x}\Big) \quad (3)$$

$$= \frac{1}{n}\frac{1 - \xi_{\boldsymbol{x}}^{n/(n-1)}}{1 - \xi_{\boldsymbol{x}}^{1/(n-1)}} \quad (4)$$

with proofs in "Methods".

We model the joint distribution of $X_1, X_2, \ldots X_d$ using a latent Gaussian copula[43]. Copulas have been used to study a wide range of dependence structures in finance[44], geology[45], and biomedicine[46] and allow us to model the density of $X$ by specifying separately the marginal distributions, easy to infer from limited samples, and the dependency structure. For a large sample space $\mathcal{X}$ and a small number $n_{\mathcal{D}}$ of available records, Gaussian copulas provide a good approximation of the density using only $d(d-1)/2$ parameters for the dependency structure and no hyperparameter.

The density of a Gaussian copula $C_\Sigma$ is expressed as:

$$c_\Sigma(\boldsymbol{u}) = \frac{1}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}\Phi^{-1}(\boldsymbol{u})^T \cdot (\Sigma^{-1} - \mathrm{I}) \cdot \Phi^{-1}(\boldsymbol{u})\right) \quad (5)$$

with a covariance matrix $\Sigma$, $\boldsymbol{u} \in [0, 1]^d$, and $\Phi$ the cumulative distribution function (CDF) of a standard univariate normal distribution.

We estimate from $\mathcal{D}$ the marginal distributions $\Psi$ (marginal parameters) for $X_1, \ldots, X_d$ and the copula distribution $\Sigma$ (covariance matrix), such that $p(\boldsymbol{x})$ is modeled by

$$q(\boldsymbol{x}|\Sigma, \Psi) = \int_{F_1^{-1}(x_1-1|\Psi)}^{F_1^{-1}(x_1|\Psi)} \cdots \int_{F_d^{-1}(x_d-1|\Psi)}^{F_d^{-1}(x_d|\Psi)} c_\Sigma(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \quad (6)$$

with $F_j$ the CDF of the discrete variable $X_j$. In practice, the copula distribution is a continuous distribution on the unit cube, and $p(\boldsymbol{x})$ its discrete counterpart on $\mathcal{X}$ (see Supplementary Methods).

We select, using maximum likelihood estimation, the marginal distributions from categorical, logarithmic, and negative binomial count distributions (see Supplementary Methods). Sampling the complete set of covariance matrices to estimate the association structure of copulas is computationally expensive for large datasets. We rely instead on a fast two-step approximate inference method: we infer separately each pairwise correlation factor $\Sigma_{ij}$ and then project the constructed matrix $\Sigma$ on the set of symmetric positive definite matrices to accurately recover the copula covariance matrix (see "Methods").

We collect five corpora from publicly available sources: population census (USA and MERNIS) as well as surveys from the UCI Machine Learning repository (ADULT, MIDUS, HDV). From each corpus, we create populations by selecting subsets of attributes (columns) uniformly. The resulting 210 populations cover a large range of uniqueness values (0–0.96), numbers of attributes (2–47), and records (7108–9M individuals). For readability purposes, we report in the main text the numerical results for all five corpora but will show figures only for USA. Figures for MERNIS, ADULT, MIDUS, and HDV are similar and available in Supplementary Information.

Figure 1a shows that, when trained on the entire population, our model correctly estimates population uniqueness $\Xi_X = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})(1 - p(\boldsymbol{x}))^{n-1}$, i.e., the expected percentage of unique individuals in $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)})$. The MAE between the empirical uniqueness of our population $\Xi_X$ and the estimated uniqueness $\widehat{\Xi_X}$ is 0.028 ± 0.026 [mean ± s.d.] for USA and 0.018 ± 0.019 on average across every corpus (see Table 1). Figure 1a and Supplementary Fig. 1 furthermore show that our model correctly estimates uniqueness across all values of uniqueness, with low within-population s.d. (Supplementary Table 3).

Figure 1b shows that our model estimates population uniqueness very well even when the dataset is heavily sampled (see Supplementary Fig. 2, for other populations). For instance, our model achieves an MAE of 0.029 ± 0.015 when the dataset only contains 1% of the USA population and an MAE of 0.041 ± 0.053 on average across every corpus. Table 1 shows that our model reaches a similarly low MAE, usually <0.050, across corpora and sampling fractions.

**Likelihood of successful re-identification**. Once trained, we can use our model to estimate the likelihood of his employer having correctly re-identified John Doe, our 50-year-old male from Berkeley with breast cancer. More specifically, given an individual record $x$, we can use the trained model to compute the likelihood $\widehat{\xi_{\boldsymbol{x}}} = (1 - q(\boldsymbol{x} \mid \Sigma, \Psi))^{n-1}$ for this record $x$ to be unique in the population. Our model takes into account information on both marginal prevalence (e.g., breast cancer prevalence) and global attribute association (e.g., gender and breast cancer). Since the cdf. of a Gaussian copula distribution has no close-form expression, we evaluate $q(\boldsymbol{x}|\Sigma, \Psi)$ with a numerical integration of the latent continuous joint density inside the hyper-rectangle defined by the $d$ components $(x_1, x_2, \ldots, x_d)$[47,48]. We assume no prior knowledge on the order of outcomes inside marginals for nominal attributes and randomize their order.

Figure 2a shows that, when trained on 1% of the USA populations, our model predicts very well individual uniqueness, achieving a mean AUC (area under the receiver-operator characteristic curve (ROC)) of 0.89. For each population, to avoid overfitting, we train the model on a single 1% sample, then select 1000 records, independent from the training sample, to test the model. For re-identifications that the model predicts to be always correct ($\widehat{\xi_{\boldsymbol{x}}} > 0.95$, estimated individual uniqueness >95%), the likelihood of them to be incorrect (false-discovery rate) is 5.26% (see bottom-right inset in Fig. 2a). ROC curves for the other populations are available in Supplementary Fig. 3 and have overall a mean AUC of 0.93 and mean false-discovery rate of 6.67% for $\widehat{\xi_{\boldsymbol{x}}} > 0.95$ (see Supplementary Table 1).

Finally, Fig. 2b shows that our model outperforms even the best theoretically achievable prediction using only population uniqueness, i.e., assigning the score $\xi_{\boldsymbol{x}}^{(\mathrm{pop})} = \Xi_X$ to every individual (ground truth population uniqueness, see Supplementary Methods). We use the Brier Score (BS)[49] to measure the calibration of probabilistic predictions: $\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}\left(\xi_{\boldsymbol{x}^{(i)}} - \widehat{\xi_{\boldsymbol{x}^{(i)}}}\right)^2$ with, in our case, $\xi_{\boldsymbol{x}^{(i)}}$ the actual uniqueness of the record $\boldsymbol{x}^{(i)}$ (1 if $\boldsymbol{x}^{(i)}$ is unique and 0 if not) and $\widehat{\xi_{\boldsymbol{x}^{(i)}}}$ the estimated likelihood. Our model obtains scores on average 39% lower than the best theoretically achievable prediction using only population uniqueness, emphasizing the importance of modeling individuals' characteristics.

**Appropriateness of the de-identification model**. Using our model, we revisit the (successful) re-identification of Gov. Weld[25]. We train our model on the 5% Public Use Microdata Sample (PUMS) files using ZIP code, date of birth, and gender and validate it using the last national estimate[50]. We show that, as a male born on July 31, 1945 and living in Cambridge (02138), the information used by Latanya Sweeney at the time, William Weld was unique with a 58% likelihood ($\xi_{\boldsymbol{x}} = 0.58$ and $\kappa_{\boldsymbol{x}} = 0.77$), meaning that Latanya Sweeney's re-identification had 77%

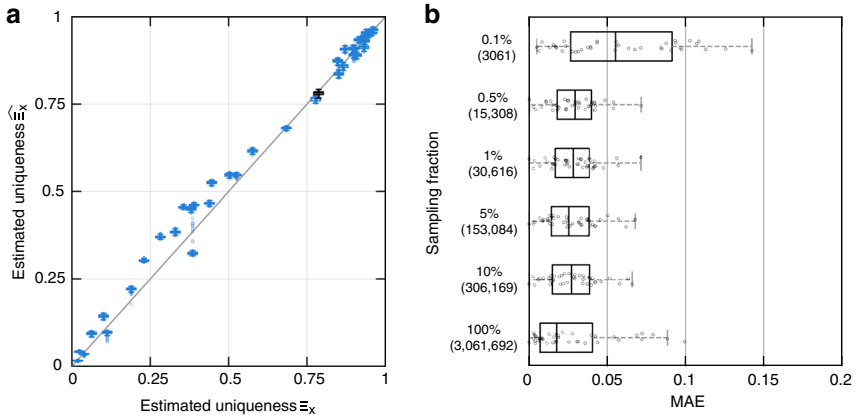


**Fig. 1** Estimating the population uniqueness of the USA corpus. **a** We compare, for each population, empirical and estimated population uniqueness (boxplot with median, 25th and 75th percentiles, maximum 1.5 interquartile range (IQR) for each population, with 100 independent trials per population). For example, date of birth, location (PUMA code), marital status, and gender uniquely identify 78.7% of the 3 million people in this population (empirical uniqueness) that our model estimates to be 78.2 ± 0.5% (boxplot in black). **b** Absolute error when estimating USA's population uniqueness when the disclosed dataset is randomly sampled from 10% to 0.1%. The boxplots (25, 50, and 75th percentiles, 1.5 IQR) show the distribution of mean absolute error (MAE) for population uniqueness, at one subsampling fraction across all USA populations (100 trials per population and sampling fraction). The *y* axis shows both *p*, the sampling fraction, and $n_S = p \times n$, the sample size. Our model estimates population uniqueness very well for all sampling fractions with the MAE slightly increasing when only a very small number of records are available ($p = 0.1\%$ or 3061 records)

**Table 1 Mean absolute error (mean ± s.d.) when estimating population uniqueness (100 trials per population)**

| | | MERNIS | USA | ADULT | HDV | MIDUS |
|---|---|---|---|---|---|---|
| Corpus | *n* | 8,820,049 | 3,061,692 | 32,561 | 8403 | 7108 |
| | *c* | 10 | 40 | 50 | 50 | 60 |
| | [min Ξ, max Ξ] | [0.087, 0.844] | [0.000, 0.961] | [0.000, 0.794] | [0.002, 0.941] | [0.052, 0.944] |
| Sampling fraction | 100% | 0.029 ± 0.019 | 0.028 ± 0.026 | 0.018 ± 0.016 | 0.006 ± 0.009 | 0.018 ± 0.014 |
| | 10% | 0.030 ± 0.019 | 0.028 ± 0.016 | 0.022 ± 0.020 | 0.011 ± 0.009 | 0.035 ± 0.044 |
| | 5% | 0.029 ± 0.019 | 0.027 ± 0.016 | 0.027 ± 0.023 | 0.015 ± 0.012 | 0.037 ± 0.055 |
| | 1% | 0.029 ± 0.019 | 0.029 ± 0.015 | 0.027 ± 0.014 | 0.045 ± 0.050 | 0.055 ± 0.079 |
| | 0.5% | 0.028 ± 0.019 | 0.029 ± 0.015 | 0.048 ± 0.039 | | |
| | 0.1% | 0.026 ± 0.017 | 0.058 ± 0.037 | | | |

Our model correctly estimates population uniqueness even when only a small to very small fraction of the population is available. *n* denotes the population size and *c* the corpus size (the total number of populations considered per corpus). We do not estimate population uniqueness when the sampled dataset contains <50 records

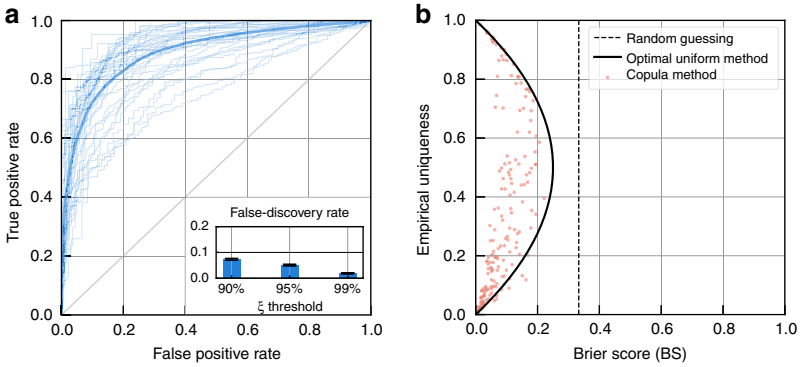


**Fig. 2** The model predicts correct re-identifications with high confidence. **a** Receiver operating characteristic (ROC) curves for USA populations (light ROC curve for each population and a solid line for the average ROC curve). Our method accurately predicts the (binary) individual uniqueness. (Inset) False-discovery rate (FDR) for individual records classified with $\xi > 0.9$, $\xi > 0.95$, and $\xi > 0.99$. For re-identifications that the model predicts are likely to be correct ($\widehat{\xi}_x > 0.95$), only 5.26% of them are incorrect (FDR). **b** Our model outperforms by 39% the best theoretically achievable prediction using population uniqueness across every corpus. A red point shows the Brier Score obtained by our model, when trained on a 1% sample. The solid line represents the lowest Brier Score achievable when using the exact population uniqueness while the dashed line represents the Brier Score of a random guess prediction (BS = 1/3)

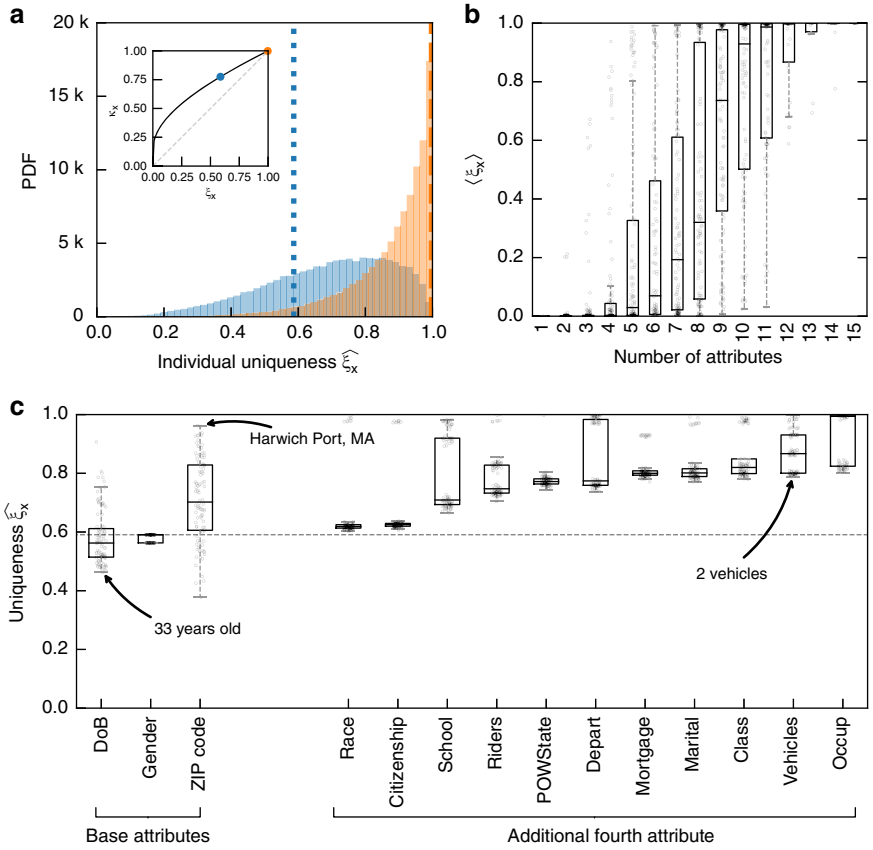


**Fig. 3** Average individual uniqueness increases fast with the number of collected demographic attributes. **a** Distribution of predicted individual uniqueness knowing ZIP code, date of birth, and gender (resp. ZIP code, date of birth, gender, and number of children) in blue (resp. orange). The dotted blue line at $\hat{\xi}_x = 0.580$ (resp. dashed orange line at $\hat{\xi}_x = 0.997$) illustrates the predicted individual uniqueness of Gov. Weld knowing the same combination of attributes. (Inset) The correctness $\kappa_x$ is solely determined by uniqueness $\xi_x$ and population size $n$ (here for Massachusetts). We show individual uniqueness and correctness for William Weld with three (in blue) and four (in orange) attributes. **b** The boxplots (25, 50, and 75th percentiles, 1.5 IQR) show the average uniqueness $\langle \xi_x \rangle$ knowing $k$ demographic attributes, grouped by number of attributes. The individual uniqueness scores $\xi_x$ are estimated on the complete population in Massachusetts, based on the 5% Public Use Microdata Sample files. While few attributes might not be sufficient for a re-identification to be correct, collecting a few more attributes will quickly render the re-identification very likely to be successful. For instance, 15 demographic attributes would render 99.98% of people in Massachusetts unique. **c** Uniqueness varies with the specific value of attributes. For instance, a 33-year-old is less unique than a 58-year-old person. We here either (*i*) randomly replace the value of one baseline attribute (ZIP code, date of birth, or gender) or (*ii*) add one extra attribute, both by sampling from its marginal distribution, to the uniqueness of a 58-year-old male from Cambridge, MA. The dashed baseline shows his original uniqueness $\hat{\xi}_x = 0.580$ and the boxplots the distribution of individual uniqueness obtained after randomly replacing or adding one attribute. A complete description of the attributes and method is available in Supplementary Methods

chances of being correct. We show that, if his medical records had included number of children—5 for William Weld—, her re-identification would have had 99.8% chances of being correct! Figure 3a shows that the same combinations of attributes (ZIP code, date of birth, gender, and number of children) would also identify 79.4% of the population in Massachusetts with high confidence ($\hat{\xi}_x > 0.80$). We finally evaluate the impact of specific attributes on William Weld's uniqueness. We either change the value of one of his baseline attributes (ZIP code, date of birth, or gender) or add one extra attribute, in both cases picking the attribute at random from its distribution (see Supplementary Methods). Figure 3c shows, for instance, that individuals with 3 cars or no car are harder to re-identify than those with 2 cars. Similarly, it shows that it would not take much to re-identify people living in Harwich Port, MA, a city of <2000 inhabitants.

Modern datasets contain a large number of points per individuals. For instance, the data broker Experian sold Alteryx access to a de-identified dataset containing 248 attributes per household for 120M Americans[51]; Cambridge university researchers shared anonymous Facebook data for 3M users collected through the myPersonality app and containing, among other attributes, users' age, gender, location, status updates, and results on a personality quiz[52]. These datasets do not necessarily share all the characteristics of the one studied here. Yet, our analysis of the re-identification of Gov. Weld by Latanya Sweeney shows that few attributes are often enough to render the likelihood of correct re-identification very high. For instance, Fig. 3b shows that the average individual uniqueness increases fast with the number of collected demographic attributes and that 15 demographic attributes would render 99.98% of people in Massachusetts unique.

Our results, first, show that few attributes are often sufficient to re-identify with high confidence individuals in heavily incomplete datasets and, second, reject the claim that sampling or releasing partial datasets, e.g., from one hospital network or a single online service, provide plausible deniability. Finally, they show that, third, even if population uniqueness is low—an argument often used to justify that data are sufficiently de-identified to be

considered anonymous[53]—, many individuals are still at risk of being successfully re-identified by an attacker using our model.

As standards for anonymization are being redefined, incl. by national and regional data protection authorities in the EU, it is essential for them to be robust and account for new threats like the one we present in this paper. They need to take into account the individual risk of re-identification and the lack of plausible deniability—even if the dataset is incomplete—, as well as legally recognize the broad range of provable privacy-enhancing systems and security measures that would allow data to be used while effectively preserving people's privacy[54,55].

## Discussion

In this paper, we proposed and validated a statistical model to quantify the likelihood for a re-identification attempt to be successful, even if the disclosed dataset is heavily incomplete.

Beyond the claim that the incompleteness of the dataset provides plausible deniability, our method also challenges claims that a low population uniqueness is sufficient to protect people's privacy[53,56]. Indeed, an attacker can, using our model, correctly re-identify an individual with high likelihood even if the population uniqueness is low (Fig. 3a). While more advanced guarantees like $k$-anonymity[57] would give every individual in the dataset some protection, they have been shown to be NP-Hard[58], hard to achieve in modern high-dimensional datasets[59], and not always sufficient[60].

While developed to estimate the likelihood of a specific re-identification to be successful, our model can also be used to estimate population uniqueness. We show in Supplementary Note 1 that, while not its primary goal, our model performs consistently better than existing methods to estimate population uniqueness on all five corpora (Supplementary Fig. 4, $P < 0.05$ in 78 cases out of 80 using Wilcoxon's signed-rank test)[61–66] and consistently better than previous attempts to estimate individual uniqueness[67,68]. Existing approaches, indeed, exhibit unpredictably large over- and under-estimation errors. Finally, a recent work quantifies the correctness of individual re-identification in incomplete (10%) hospital data using complete population frequencies[24]. Compared to this work, our approach does not require external data nor to assume this external data to be complete.

To study the stability and robustness of our estimations, we perform further experiments (Supplementary Notes 2–8).

First, we analyze the impact of marginal and association parameters on the model error and show how to use exogenous information to lower it. Table 1 and Supplementary Note 7 show that, at very small sampling fraction (below 0.1%), where the error is the largest, the error is mostly determined by the marginals, and converges after few hundred records when the exact marginals are known. The copula covariance parameters exhibit no significant bias and decrease fast when the sample size increases (Supplementary Note 8).

As our method separates marginals and association structure inference, exogenous information from larger data sources could also be used to estimate marginals with higher accuracy. For instance, count distributions for attributes such as date of birth or ZIP code could be directly estimated from national surveys. We replicate our analysis on the USA corpus using a subsampled dataset to infer the association structure along with the exact counts for marginal distributions. Incorporating exogenous information reduces, e.g., the mean MAE of uniqueness across all corpora by 48.6% ($P < 0.01$, Mann–Whitney) for a 0.1% sample. Exogenous information become less useful as the sampling fraction increases (Supplementary Table 2).

Second, our model assumes that $\mathcal{D}$ is either uniformly sampled from the population of interest $X$ or, as several census bureaus are doing, released with post-stratification weights to match the overall population. We believe this to be a reasonable assumption as biases in the data would greatly affect its usefulness and affect any application of the data, including our model. To overcome an existing sampling bias, the model can be (*i*) further trained on a random sample from the population $\mathcal{D}$ (e.g., microdata census or survey data) and then applied to a non-uniform released sample (e.g., hospital data, not uniformly sampled from the population) or (*ii*) trained using better, potentially unbiased, estimates for marginals or association structure coming from other sources (see above).

Third, since $\mathcal{D}$ is a sample from the population $X$, only the records that are unique in the sample can be unique in the population. Hence, we further evaluate the performance on our model only on records that are sample unique and show that it only marginally decrease the AUC (Supplementary Note 5). We therefore prefer to not restrict our predictions to sample unique records as (a) our models need to perform well on non-sample unique records for us to be able to estimate correctness and (b) to keep the method robust if oversampling or sampling with replacement were to have been used.

## Methods

**Inferring marginals distributions**. Marginals can be either (i) unknown and are estimated from the marginals of the population sample $X_S$, this is the assumption used in the main text, or (ii) known with their exact distribution and cumulative density function directly available.

In the first case, we fit marginal counts to categorical (naive plug-in estimator), negative binomial, and logarithmic distributions using maximum log-likelihood. We compare the obtained distributions and select the best likelihood according to its Bayesian information criterion (BIC):

$$\mathrm{BIC} = -2\log\widehat{L} + k\log n_{\mathcal{D}} \tag{7}$$

where $\widehat{L}$ is the maximized value of the likelihood function, $n_{\mathcal{D}}$ the number of individuals in the sample $\mathcal{D}$, and $k$ the number of parameters in the fitted marginal distribution.

**Inferring the parameters of the latent copula**. Each cell $\Sigma_{ij}$ of the $\Sigma$ covariance matrix of a multivariate copula distribution is the correlation parameter of a pairwise copula distribution. Hence, instead of inferring $\Sigma$ from the set of all covariance matrices, we separately infer every cell $\Sigma_{ij} \in [0, 1]$ from the joint sample of $\mathcal{D}_i$ and $\mathcal{D}_j$. We first measure the mutual information $I(\mathcal{D}_i; \mathcal{D}_j)$ between the two attributes and select $\sigma = \widehat{\Sigma_{ij}}$ minimizing the Euclidean distance between the empirical mutual information and the mutual information of the inferred joint distribution.

In practice, since the cdf. of a Gaussian copula is not tractable, we use a bounded Nelder–Mead minimization algorithm. For a given $(\sigma, (\Psi_i, \Psi_j))$, we sample from the distribution $q(\cdot|\sigma, (\Psi_i, \Psi_j))$ and generate a discrete bivariate sample $Y$ from which we measure the objective:

$$f(\sigma) = \begin{cases} \left\| I(\mathcal{D}_i; \mathcal{D}_j) - I(Y_1; Y_2) \right\|_2 & \text{for } \sigma \in [0, 1] \\ +\infty & \text{otherwise} \end{cases} \tag{8}$$

We then project the obtained $\widehat{\Sigma}$ matrix on the set of SDP matrices by solving the following optimization problem:

$$\begin{aligned} \min_{A} \quad & \left\| A - \widehat{\Sigma} \right\|_2 \\ \text{s.t.} \quad & A \succcurlyeq 0 \end{aligned} \tag{9}$$

**Modeling the association structure using mutual information**. We use the pairwise mutual information to measure the strength of association between attributes. For a dataset $\mathcal{D}$, we denote by $I_{\mathcal{D}}$ the mutual information matrix where each cell $I(\mathcal{D}_i; \mathcal{D}_j)$ is the mutual information between attributes $\mathcal{D}_i$ and $\mathcal{D}_j$. When evaluating mutual information from small samples, obtained scores are often overestimating the strength of association. We apply a correction for randomness using a permutation model[69]:

$$AI(\mathcal{D}_i; \mathcal{D}_j) = \frac{I(\mathcal{D}_i; \mathcal{D}_j) - \mathbb{E}(I(\mathcal{D}_i; \mathcal{D}_j))}{\max\{\mathbb{H}(\mathcal{D}_i), \mathbb{H}(\mathcal{D}_j)\} - \mathbb{E}(I(\mathcal{D}_i; \mathcal{D}_j))} \tag{10}$$

In practice, we estimate the expected mutual information between $\mathcal{D}_i$ and $\mathcal{D}_j$ with successive permutations of $\mathcal{D}_j$. We found that the adjusted mutual information provides significant improvement for small samples and large support size $|\mathcal{X}|$ compared to the naive estimator.

**Theoretical and empirical population uniqueness**. For $n$ individuals $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$, …, $\boldsymbol{x}^{(n)}$ drawn from $X$, the uniqueness $\Xi_X$ is the expected percentage of unique individuals. It can be estimated either (i) by computing the mean of individual uniqueness or (ii) by sampling a synthetic population of $n$ individuals from the copula distribution. In the former case, we have

$$\Xi_X \equiv \frac{1}{n}\,\mathbb{E}\left[\sum_{i=1}^{n}\left[\boldsymbol{x}^{(i)}\text{unique in}(\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(n)})\right]\right] \tag{11}$$

$$= \frac{1}{n}\,\mathbb{E}\left[\sum_{\boldsymbol{x}\in\mathcal{X}} T_{\boldsymbol{x}}\right] \tag{12}$$

$$= \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{X}}\mathbb{E}[T_{\boldsymbol{x}}] \tag{13}$$

where $T_{\boldsymbol{x}} = [\exists! i, \boldsymbol{x}^{(i)} = \boldsymbol{x}]$ equals one if there exists a single individual $i$ such as $\boldsymbol{x}^{(i)} = \boldsymbol{x}$ and zero otherwise. $T_{\boldsymbol{x}}$ follows a binomial distribution $B(p(\boldsymbol{x}), n)$. Therefore

$$\mathbb{E}[T_{\boldsymbol{x}}] = np(\boldsymbol{x})\,(1-p(\boldsymbol{x}))^{n-1} \tag{14}$$

and

$$\Xi_X = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})(1-p(\boldsymbol{x}))^{n-1} \tag{15}$$

This requires iterating over all combinations of attributes, whose number grows exponentially as the number of attributes increases, and quickly becomes computationally intractable. The second method is therefore often more tractable and we use it to estimate population uniqueness in the paper.

For cumulative marginal distributions $F_1$, $F_2$, …, $F_d$ and copula correlation matrix $\Sigma$, the algorithm 1 (Supplementary Methods) samples $n$ individuals from $q(\cdot|\Sigma,\Psi)$ using the latent copula distribution. From the $n$ generated records ($\boldsymbol{y}^{(1)}$, $\boldsymbol{y}^{(2)}$, …, $\boldsymbol{y}^{(n)}$), we compute the empirical uniqueness

$$\Xi_X = \frac{1}{n}\left|\left\{i \in [1,n] \;/\; \forall j\neq i, \boldsymbol{y}^{(i)}\neq\boldsymbol{y}^{(j)}\right\}\right| \tag{16}$$

**Individual likelihood of uniqueness and correctness**. The probability distribution $q(\cdot\,|\,\Sigma,\Psi)$ can be computed by integrating over the latent copula density. Note that the marginal distributions $X_1$ to $X_d$ are discrete, causing the inverses $F_1^{-1}$ to $F_d^{-1}$ to have plateaus. When estimating $p(\boldsymbol{x})$, we integrate over the latent copula distribution inside the hypercube $[x_1-1, x_1]\times[x_2-1, x_2]\times\ldots\times[x_d-1, x_d]$:

$$q(\boldsymbol{x}\,|\Sigma,\Psi) = \mathbb{P}(x_1-1<X_1\le x_1,\ldots,x_d-1<X_d\le x_d|\Sigma,\Psi) \tag{17}$$

$$= \int_{F_1^{-1}(x_1-1|\Psi)}^{F_1^{-1}(x_1|\Psi)}\cdots\int_{F_d^{-1}(x_d-1|\Psi)}^{F_d^{-1}(x_d|\Psi)} c_\Sigma(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \tag{18}$$

$$= \int_{\phi^{-1}(F_1^{-1}(x_1-1|\Psi))}^{\phi^{-1}(F_1^{-1}(x_1|\Psi))}\cdots\int_{\phi^{-1}(F_d^{-1}(x_d-1|\Psi))}^{\phi^{-1}(F_d^{-1}(x_d|\Psi))} \phi_\Sigma(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \tag{19}$$

with $\phi_\Sigma$ the density of a zero-mean multivariate normal (MVN) of correlation matrix $\Sigma$. Several methods have been proposed in the literature to estimate MVN rectangle probabilities. Genz and Bretz[47,48] proposed a randomized quasi Monte Carlo method which we use to estimate the discrete copula density.

The likelihood $\xi_{\boldsymbol{x}}$ for an individual's record $\boldsymbol{x}$ to be unique in a population of $n$ individuals can be derived from $p_X(X=\boldsymbol{x})$:

$$\xi_{\boldsymbol{x}} \equiv p_X(\boldsymbol{x}\ \text{unique in}\ (\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(n)})\mid\exists i,\boldsymbol{x}^{(i)}=\boldsymbol{x}) \tag{20}$$

$$= p_X(\boldsymbol{x}\ \text{unique in}\ (\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(n)})\mid\boldsymbol{x}^{(1)}=\boldsymbol{x}) \tag{21}$$

$$= p_X(\forall i\in[2,n],\boldsymbol{x}^{(i)}\neq\boldsymbol{x}) \tag{22}$$

$$= (1-p(\boldsymbol{x}))^{n-1} \tag{23}$$

$$\widehat{\xi}_{\boldsymbol{x}} = (1-q(\boldsymbol{x}\,|\,\Sigma,\Psi))^{n-1}$$

Similarly, the likelihood $\kappa_{\boldsymbol{x}}$ for an individual's record $x$ to be correctly matched in a population of $n$ individuals can be derived from $p_X(X=\boldsymbol{x})$. With $T\equiv\sum_{i=1}^{n}[\boldsymbol{x}^{(i)}=\boldsymbol{x}]-1$, the number of potential false positives in the population, we have:

$$\kappa_{\boldsymbol{x}} \equiv \mathbb{P}(\boldsymbol{x}\ \text{correctly matched in}\ (\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(n)})\mid\exists i,\boldsymbol{x}^{(i)}=\boldsymbol{x}) \tag{24}$$

$$= \sum_{k=0}^{n-1}\frac{1}{k+1}\mathbb{P}(T=k) \tag{25}$$

$$= \sum_{k=0}^{n-1}\frac{1}{k+1}\binom{n-1}{k}p(\boldsymbol{x})^k(1-p(\boldsymbol{x}))^{(n-1-k)} \tag{26}$$

$$= \frac{1}{np(\boldsymbol{x})}(1-(1-p(\boldsymbol{x}))^n) \tag{27}$$

Note that, since records are independent, $T$ follows a binomial distribution $B(n-1, p(\boldsymbol{x}))$.

We substitute the expression for $\xi_x$ in the last formula and obtain:

$$\kappa_{\boldsymbol{x}} = \frac{1}{np(\boldsymbol{x})}(1-(1-p(\boldsymbol{x}))^n) \tag{28}$$

$$= \frac{1}{n}\frac{1-\xi_{\boldsymbol{x}}^{n/(n-1)}}{1-\xi_{\boldsymbol{x}}^{1/(n-1)}} \tag{29}$$

## Data availability

The USA corpus, extracted from the 1-Percent Public Use Microdata Sample (PUMS) files, is available at https://www.census.gov/main/www/pums.html. The 5% PUMS files used to estimate the correctness of Governor Weld's re-identification are also available at the same address. The ADULT corpus, extracted from the Adult Income dataset, is available at https://archive.ics.uci.edu/ml/datasets/adult. The HDV corpus, extracted from the Histoire de vie survey, is available at https://www.insee.fr/fr/statistiques/2532244. The MIDUS corpus, extracted from the Midlife in the United States survey, is available at https://www.icpsr.umich.edu/icpsrweb/ICPSR/series/203. The MERNIS corpus is extracted from a complete population database of virtually all 48 million individuals born before early 1991 in Turkey that was made available online in April 2016 after a data leak from Turkey's Central Civil Registration System. Our use of this data was approved by Imperial College as it provides a unique opportunity to perform uniqueness estimation on a complete census survey. Owing to the sensitivity of the data, we have only analyzed a copy of the dataset where every distinct value was replaced by a unique integer to obfuscate records, without loss of precision for uniqueness modeling. A complete description of each corpus is available in the Supplementary Information.

## Code availability

All simulations were implemented in Julia and Python. The source code to reproduce the experiments is available at https://cpg.doc.ic.ac.uk/individual-risk, along with documentation, tests, and examples.

Received: 27 September 2018 Accepted: 11 June 2019
Published online: 23 July 2019

## References

1. Poushter, J. Smartphone ownership and internet usage continues to climb in emerging economies (Pew Research Center, Washington, DC, 2016). http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/.
2. Yang, N. & Hing, E. National electronic health records survey. https://cdc.gov/nchs/data/ahcd/nehrs/2015_nehrs_ehr_by_specialty.pdf (2015).
3. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *JAMA* **309**, 1351–1352 (2013).
4. Wyber, R. et al. Big data in global health: improving health in low- and middle-income countries. *Bull. World Health Organ.* **93**, 203–208 (2015).
5. Lazer, D. et al. Life in the network: the coming age of computational social science. *Science* **323**, 721 (2009).
6. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).
7. Kitchin, R. The real-time city? Big data and smart urbanism. *GeoJournal* **79**, 1–14 (2014).
8. McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J. & Barton, D. Big data: the management revolution. *Harv. Bus. Rev.* **90**, 60–68 (2012).
9. Hodson, H. Revealed: Google AI has access to huge haul of NHS patient data. *New Scientist* (29 Apr 2016).
10. Cadwalladr, C. & Graham-Harrison, E. Revealed: 50 million facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian* (17 Mar 2018).
11. Morey, T., Forbath, T. & Schoop, A. Customer data: designing for transparency and trust. *Harv. Bus. Rev.* **93**, 96–105 (2015).

12. Polonetsky, J., Tene, O. & Finch, K. Shades of gray: seeing the full spectrum of practical data De-Identification. *Santa Clara Law Rev.* **56**, 593–629 (2016).
13. Office for Civil Rights, HHS. *Standards for privacy of individually identifiable health information*. Federal Register. https://ncbi.nlm.nih.gov/pubmed/12180470 (2002).
14. Malin, B., Benitez, K. & Masys, D. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA privacy rule. *J. Am. Med. Inform. Assoc.* **18**, 3–10 (2011).
15. Rothstein, M. A. Is deidentification sufficient to protect health privacy in research? *Am. J. Bioeth.* **10**, 3–11 (2010).
16. Council of European Union. Regulation (EU) 2016/679. *Off. J. Eur. Union* **L 119**, 1–88 (2016).
17. Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J. & Altman, D. G. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *BMJ* **340**, c181 (2010).
18. Opinion 05/2014 on anonymisation techniques. Technical Report, Article 29 Data Protection Working Party. http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (2014).
19. Rubinstein, I. Framing the discussion. https://fpf.org/wp-content/uploads/2016/11/Rubinstein_framing-paper.pdf (2016).
20. Cal. Civil Code. Assembly Bill No. 375 §§ 1798.100–1798.198 (2018).
21. Narayanan, A. & Felten, E. W. No silver bullet: de-identification still doesn't work. http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf (2014).
22. Ohm, P. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA. Law Rev.* **57**, 1701 (2010).
23. Hern, A. 'Anonymous' browsing data can be easily exposed, researchers reveal. *The Guardian* (1 Aug 2017).
24. Culnane, C., Rubinstein, B. I. P. & Teague, V. Health data in an open world. Preprint at: https://arxiv.org/abs/1712.05627 (2017).
25. Sweeney, L. Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics* **25**, 98–110. 82 (1997).
26. Loukides, G., Denny, J. C. & Malin, B. The disclosure of diagnosis codes can breach research participants' privacy. *J. Am. Med. Inform. Assoc.* **17**, 322–327 (2010).
27. Douriez, M., Doraiswamy, H., Freire, J. & Silva, C. T. Anonymizing NYC taxi data: does it matter? In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 140–148 (IEEE, Piscataway, NJ, 2016).
28. Siddle, J. I know where you were last summer: London's public bike data is telling everyone where you've been. https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html (2014). Accessed 7 Feb 2019.
29. Lavrenovs, A. & Podins, K. Privacy violations in Riga open data public transport system. In *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, 1–6 (IEEE, Piscataway, NJ, 2016). https://doi.org/10.1109/AIEEE.2016.7821808.
30. de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. & Blondel, V. D. Unique in the crowd: the privacy bounds of human mobility. *Sci. Rep.* **3**, 1376 (2013).
31. de Montjoye, Y.-A., Radaelli, L., Singh, V. K. & Pentland, A. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science* **347**, 536–539 (2015).
32. Matthews, G. J. & Harel, O. Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy. *Stat. Surv.* **5**, 1–29 (2011).
33. Barth-Jones, D. The 're-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. https://ssrn.com/abstract=2076397 (2012).
34. El Emam, K. & Arbuckle, L. De-identification: a critical debate. https://fpf.org/2014/07/24/de-identification-a-critical-debate/ (2014).
35. Sánchez, D., Martnez, S. & Domingo-Ferrer, J. Comment on "unique in the shopping mall: on the reidentifiability of credit card metadata". *Science* **351**, 1274 (2016).
36. Reiter, J. P. Estimating risks of identification disclosure in microdata. *J. Am. Stat. Assoc.* **100**, 1103–1112 (2005).
37. Fienberg, S. E. & Sanil, A. P. A Bayesian approach to data disclosure: optimal intruder behavior for continuous data. *J. Stat.* **13**, 75 (1997).
38. Duncan, G. & Lambert, D. The risk of disclosure for microdata. *J. Bus. Econ. Stat.* **7**, 207–217 (1989).
39. Office of the Australian Information Commissioner. *De-identification and the Privacy Act*. https://www.oaic.gov.au/agencies-and-organisations/guides/de-identification-and-the-privacy-act (2018).
40. Ruggles, S., King, M. L., Levison, D., McCaa, R. & Sobek, M. IPUMS-International. *Hist. Methods* **36**, 60–65 (2003).
41. Bennett, J. & Lanning, S. The Netflix prize. In *Proc. KDD Cup and Workshop*, 35–38 (ACM, New York, NY, 2007). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.6998.
42. Sweeney, L. Simple demographics often identify people uniquely. *Health* **671**, 1–34 (2000).
43. Genest, C. & Mackay, J. The joy of copulas: bivariate distributions with uniform marginals. *Am. Stat.* **40**, 280–283 (1986).
44. Cherubini, U., Luciano, E. & Vecchiato, W. *Copula Methods in Finance* (Wiley-Blackwell, Hoboken, NJ, 2004).
45. Genest, C. & Favre, A.-C. Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.* **12**, 347–368 (2007).
46. Wang, W. & Wells, M. T. Model selection and semiparametric inference for bivariate failure-time data. *J. Am. Stat. Assoc.* **95**, 62–72 (2000).
47. Genz, A. Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Stat.* **1**, 141–149 (1992).
48. Genz, A. & Bretz, F. *Computation of Multivariate Normal and t Probabilities* (Springer Science & Business Media, Berlin, 2009).
49. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
50. Golle, P. Revisiting the uniqueness of simple demographics in the US population. In *5th ACM Workshop on Privacy in Electronic Society* (ACM, New York, NY, 2006). https://doi.org/10.1145/1179601.1179615.
51. Fox-Brewster, T. 120 million american households exposed in 'massive' ConsumerView database leak. *Forbes* (2017).
52. Waterfield, P. & Revell, T. Huge new facebook data leak exposed intimate details of 3m users. *New Scientist* (2018).
53. El Emam, K. & Arbuckle, L. *Anonymizing Health Data* (O'Reilly, Newton, MA, 2013).
54. D'Acquisto, G. et al. Privacy by design in big data: an overview of privacy enhancing technologies in the era of big data analytics. Technical Report. European Union Agency for Network and Information Security (2015).
55. Cho, H., Wu, D. J. & Berger, B. Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnol.* **36**, 547–551 (2018).
56. Cavoukian, A. & Castro, D. *Big data and innovation, setting the record straight: de-identification does work*. http://www2.itif.org/2014-big-data-deidentification.pdf (2014).
57. Sweeney, L. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **10**, 557–570 (2002).
58. Meyerson, A. & Williams, R. On the complexity of optimal k-anonymity. In *Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 223–228 (2004). https://doi.org/10.1145/1055558.1055591.
59. Aggarwal, C. C. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, 901–909 (VLDB Endowment, 2005). http://dl.acm.org/citation.cfm?id=1083592.1083696.
60. Li, N., Li, T. & Venkatasubramanian, S. t-closeness: privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, 106–115 (IEEE, 2007). https://doi.org/10.1109/ICDE.2007.367856.
61. Ewens, W. J. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112 (1972).
62. Chen, G. & Keller-McNulty, S. Estimation of identification disclosure risk in microdata. *J. Stat.* **14**, 79 (1998).
63. Hoshino, N. Applying pitman's sampling formula to microdata disclosure risk assessment. *J. Stat.* **17**, 499 (2001).
64. Keller, W. J. & Pannekoek, J. Disclosure control of microdata. *J. Am. Stat. Assoc.* **85**, 38–45 (1990).
65. Dankar, F. K., El Emam, K., Neisa, A. & Roffey, T. Estimating the re-identification risk of clinical data sets. *BMC Med. Inform. Decis. Mak.* **12**, 66 (2012).
66. Pitman, J. Random discrete distributions invariant under size-biased permutation. *Adv. Appl. Probab.* **28**, 525–539 (1996).
67. Skinner, C. J. & Holmes, D. J. Estimating the re-identification risk per record in microdata. *J. Stat.* **14**, 361 (1998).
68. Skinner, C. & Shlomo, N. Assessing identification risk in survey microdata using Log-Linear models. *J. Am. Stat. Assoc.* **103**, 989–1001 (2008).
69. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010).


## Acknowledgements

L.R. is the recipient of a doctoral fellowship from the Belgian Fund for Scientific Research (F.R.S.-FNRS). This collaboration was made possible thanks to Imperial College's European Partners Fund and a WBI World Excellence Grant. We acknowledge support from the Information Commissioner Office for the development of the online demonstration tool.

## Author contributions

L.R. designed and performed experiments, analyzed the data, and wrote the paper; Y.-A. d.M. and J.M.H. designed experiments and wrote the paper.

## Additional information

**Supplementary Information** accompanies this paper at https://doi.org/10.1038/s41467-019-10933-3.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**Peer review information**: *Nature Communications* thanks Antoine Boutet, Vanessa Teague, and other anonymous reviewer(s) for their contribution to the peer review of this work.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.