Generating synthetic mixed discrete-continuous health records with mixed sum-product networks

Kroes, S.K.S.; Leeuwen, M. van; Groenwold, R.H.H.; Janssen, M.P.

Journal of the American Medical Informatics Association, 30(1), 2023, 16–25
https://doi.org/10.1093/jamia/ocac184
Advance Access Publication Date: 13 October 2022
Research and Applications

Research and Applications

Generating synthetic mixed discrete-continuous health records with mixed sum-product networks

Shannon K.S. Kroes [1,2,3], Matthijs van Leeuwen [2], Rolf H.H. Groenwold [3,4], and Mart P. Janssen [1,2]

[1]Transfusion Technology Assessment Group, Donor Medicine Research Department, Sanquin Research, Amsterdam, The Netherlands, [2]Leiden Institute of Advanced Computer Science, Computer Science, Leiden University, Leiden, The Netherlands, [3]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands, and [4]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Corresponding Author: Mart P. Janssen, PhD, Transfusion Technology Assessment Group, Donor Medicine Research Department, Sanquin Research, Plesmanlaan 125, 1066 CX Amsterdam, The Netherlands; m.janssen@sanquin.nl

Received 24 July 2022; Revised 9 September 2022; Editorial Decision 14 September 2022; Accepted 1 October 2022

ABSTRACT

Objective: Privacy is a concern whenever individual patient health data is exchanged for scientific research. We propose using mixed sum-product networks (MSPNs) as private representations of data and take samples from the network to generate synthetic data that can be shared for subsequent statistical analysis. This anonymization method was evaluated with respect to privacy and information loss.

Materials and methods: Using a simulation study, information loss was quantified by assessing whether synthetic data could reproduce regression parameters obtained from the original data. Predictors variable types were varied between continuous, count, categorical, and mixed discrete-continuous. Additionally, we measured whether the MSPN approach successfully anonymizes the data by removing associations between background and sensitive information for these datasets.

Results: The synthetic data generated with MSPNs yielded regression results highly similar to those generated with original data, differing less than 5% in most simulation scenarios. Standard errors increased compared to the original data. Particularly for smaller datasets (1000 records), this resulted in a discrepancy between the estimated and empirical standard errors. Sensitive values could no longer be inferred from background information for at least 99% of tested individuals.

Discussion: The proposed anonymization approach yields very promising results. Further research is required to evaluate its performance with other types of data and analyses, and to predict how user parameter choices affect a bias-privacy trade-off.

Conclusion: Generating synthetic data from MSPNs is a promising, easy-to-use approach for anonymization of sensitive individual health data that yields informative and private data.

Key words: privacy, anonymization, mixed sum-product networks, health data, synthetic data

INTRODUCTION

Whenever individual patient health data needs to be exchanged for scientific research, privacy is an important concern. Privacy-enhancing technologies can safeguard patient privacy while creating opportunities for scientific research and collaboration. An increasingly popular privacy-enhancing approach is to create synthetic data, which can be used in scenarios where the data holder does not have the resources or statistical expertise to carry out and design the appropriate analyses.

Synthetic data can be created by generating records from a representation of the data, for example by sampling from a model of the privacy-sensitive data. Such approaches have been proposed with generative adversarial networks,[1–3] hashing,[4] support vector machines,[5] Bayesian networks,[6] and copulas.[7] However, the extent to which these synthetic data protect privacy can be difficult to assess. Particularly when the differential privacy framework is used,[8] the user has to choose a level of privacy by setting a parameter ε, which is difficult to interpret.[9] Furthermore, most approaches are not suited for mixed discrete-continuous data, whereas clinical data typically contain a mixed set of variables.

An approach to model data that has not yet been used to generate synthetic data but has potential to overcome these limitations, is a *mixed sum-product network* (MSPN). Sum-product networks are directed acyclic graphs that can be used to describe joint probability distributions.[10,11] Recently, they have been extended to mixed discrete-continuous data (hence MSPNs) with nonparametric estimation methods and have been shown to be competitive with hybrid Bayesian networks.[12]

MSPNs are particularly suited to generate synthetic, anonymized data, because they break down associations that can facilitate the extraction of sensitive information. This is achieved by creating clusters in the data and assuming independence between variables within these clusters.

Clustering is a fundamental element of many anonymization approaches in the field of privacy-preserving data publication and sharing,[13] including generalization,[14,15] bucketization,[16] slicing,[17] and disassociation.[18] The type of clusters created with MSPNs can be seen as a form of generalization and disassociation. Generalization hinders the use of background information by categorizing values such that more overlap is created between information of different individuals. Disassociation impedes possibilities for attackers to use background information to infer sensitive information by removing their connections. Neither of these concepts have been applied to the field of "randomization," in which random noise is added to synthetic data; thus creating uncertainty for a potential attacker of sensitive data.[19]

We propose an anonymization approach where we use constrained MSPNs to remove connections in the data that might otherwise facilitate the extraction of sensitive information, and draw random samples from the MSPNs to generate synthetic data. This approach can be seen as a hybrid between randomization, disassociation, and generalization, where the MSPN serves as a private representation of the data. In this work, we studied whether MSPNs indeed succeed in removing opportunities to infer privacy-sensitive information. Furthermore, we established whether MPSNs simultaneously retain the information required to answer clinical research questions. We evaluate whether synthetic data generated with MSPNs lead to the same linear regression parameter estimates as the original, privacy-sensitive data. This frequently used and versatile method of analysis is evaluated on different types of data, including mixed discrete-continuous data.

MATERIALS AND METHODS

The anonymization approach

We used 2-layered Mixed Sum-Products Network (MSPNs) and sample from the MSPNs to generate synthetic data. A 2-layered MSPN is comprised of weights, product nodes, histograms, and one sum node. We refer to Figure 1 for an example of an MSPN. When creating a sum node, the data are partitioned into γ distinct clusters, where similar individuals are grouped. The probabilities of value combinations from within these clusters are described with weights that correspond to the cluster sizes. These weights and clusters model associations between variables. For example, if a dataset is a random sample from the variables *age*, *gender*, and *heart failure*, the MSPN could group older men with heart failure in one cluster and younger women with heart failure in another cluster. Because the first group is likely much larger than the second, this cluster has a higher weight than the second cluster, thus indicating that values in this cluster are more probable. Thereby, the MSPN represents that older age and being male are associated with a higher probability of heart failure.

Using product nodes, the probability of a combination of values *within* a cluster is computed by taking the product of the univariate probabilities for each of the variables (see Figure 1). The probability distribution of each variable is described *separately* with a histogram density estimator. As a result, within clusters, the information on which values belong to the same record is removed. For example, if the data contain a 50-year-old female and a 61-year-old male, it cannot be inferred which age and which gender belonged to the same record, provided these observations fall within the same cluster. Furthermore, the histograms categorize continuous variables and thus remove the information on which *specific* continuous values occurred in the original data. This disassociation and generalization within clusters considerably impedes an attackers' ability to extract sensitive information by using values on other (background) variables.

The sum node, weights, product nodes, and histograms describe the probability distribution. For example, in Figure 1, the combination $X_1 = 0$ and $X_2 = -2$ falls in cluster 1. Therefore, the probability of this data point is equal to $P_{1,X1}(X_1 = 0)*P_{1,X2}(X_2 = 2)*0.25$, where $P_{1,X1}$ and $P_{1,X2}$ indicate the univariate distributions of X_1 and X_2 in cluster 1, respectively (ie, the 2 histograms attached to the product node for cluster 1), and 0.25 is the proportion of individuals represented by cluster 1.

To obtain synthetic data that can be used for statistical analyses, instead of the original, privacy-sensitive data, a random sample is taken from the MSPN. This random process further increases uncertainty for an attacker, as the synthetic records do not correspond directly to the individuals in the original data. Specifics on our implementation will be provided in a later subsection.

Simulation

Figure 2 illustrates the simulation process used to evaluate whether the anonymized data can be used as a substitute for the original data in linear regression analyses. By varying parameters of the data generating mechanism of the original data, in total 24 simulation scenarios were studied with a full factorial design: 3 different sample sizes, 4 types of predictors, and either a zero or nonzero regression parameter value of the predictor of interest. For each of these scenarios, the process was repeated multiple times. Details on each of the components of the simulation are provided below.

Data generating mechanism of the original data

Let $\mathbf{X}$ denote the original dataset, consisting of 9 variables in every scenario, where X_1–X_8 serve as *predictors* and X_9 is the *outcome variable*. The number of records (n) was varied between 1000, 10 000, and 100 000. All continuous variables were rounded to 2 decimals to mimic the precision of measurements in clinical medical practice.

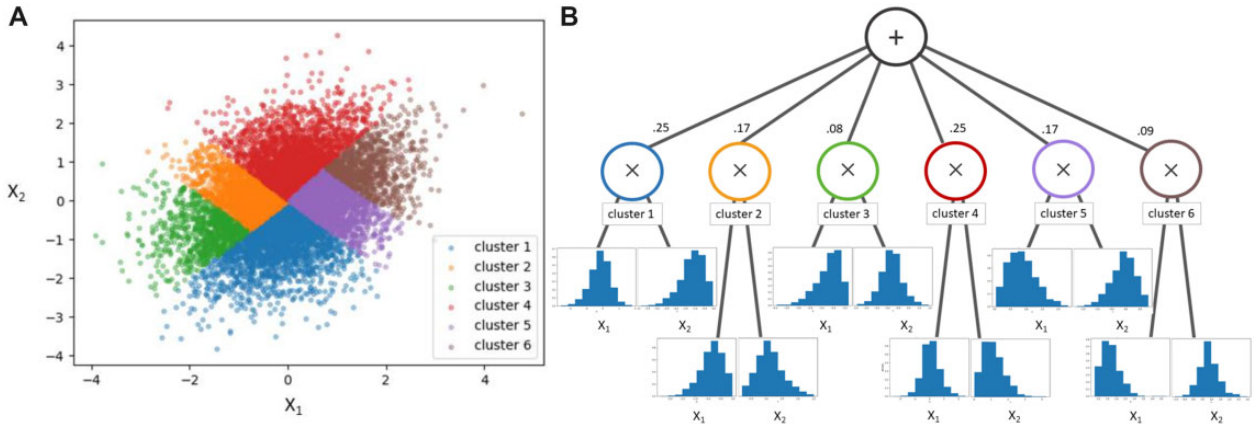

Figure 1. (A) Scatterplot of variables X_1 and X_2 of a bivariate normal distribution with correlation of 0.3 ($n = 10\,000$), where the color of each data point corresponds to the cluster to which it has been assigned. (B) The corresponding MSPN, containing one sum node and 6 product nodes, with the histograms representing the probability densities of the (independent) univariate distributions.

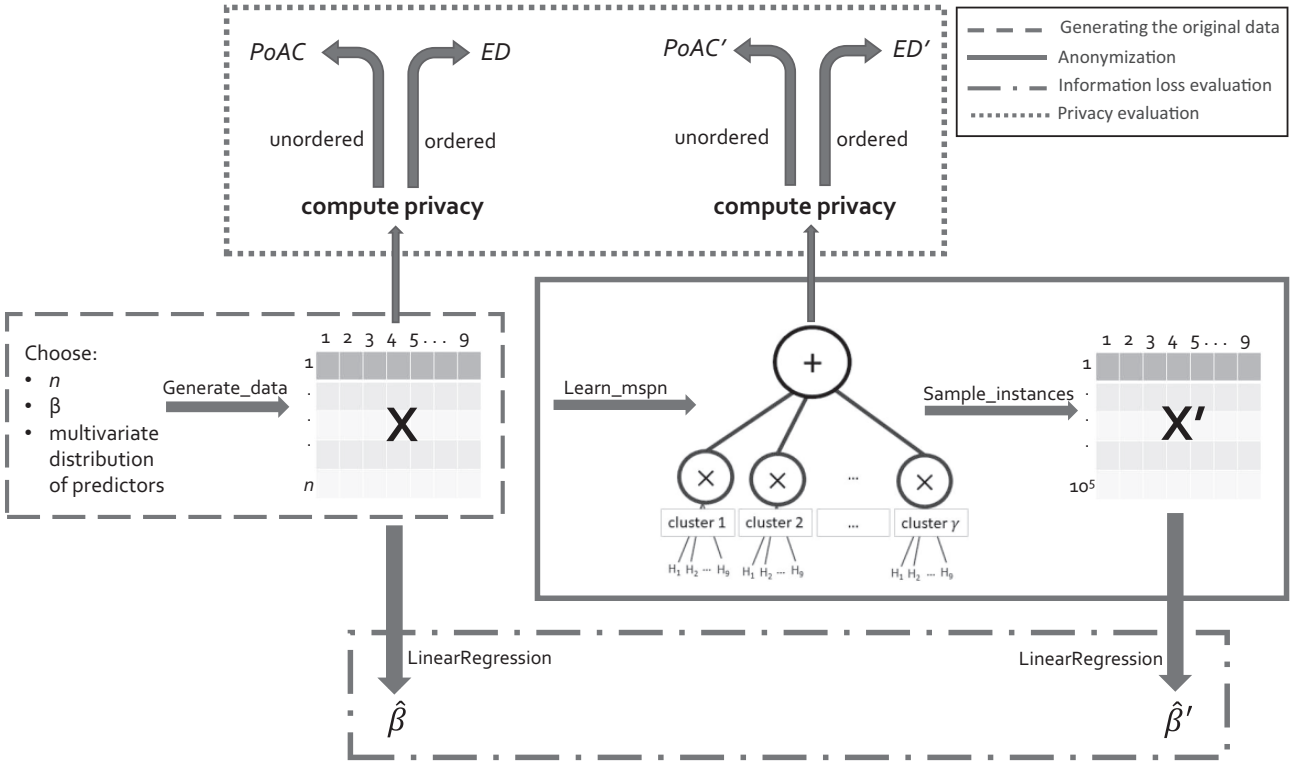

Figure 2. Illustration of the simulation study that evaluates the performance of the proposed anonymization method with MSPNs. **X** = original dataset; **X′** = corresponding synthetic dataset; n = number of records in **X** (either 1000, 10 000, or 100 000); β = true value of regression parameter of interest (either zero or non-zero); multivariate distribution of predictors is varied between continuous, count-valued, discrete, or mixed; $\hat{\beta}$ = parameter of interest as estimated with **X**; $\hat{\beta}'$ = parameter of interest as estimated with **X′**; H_j = histogram of the jth variable of **X** within cluster; γ = number of clusters; *PoAC* = Proportion of Alternatives Considered (privacy measure for unordered variables); *ED* = expected deviation (privacy measure for ordered variables). An arrow with a function name (capital first letter) points from its input to its output. *Note*: The function names in this figure correspond to the code in the GitHub repository: (https://github.com/ShannonKroes/MSPN_privacy).

Generating the predictors. To generate the values for the predictors, first n samples were generated from an 8-variate standard normal distribution, with pairwise correlations of 0.3. Next, for each sample, its exceedance probability was computed and the value corresponding to that probability was extracted from the cumulative density function of the desired univariate (parametric) distribution (eg, a Poisson distribution). This allowed the introduction of correlations between all variables irrespective of their type.

We varied the distribution of the predictors between the following options:

- Continuous: multivariate standard normal.

- Count: Poisson distributed, $\lambda_1, \lambda_2, \lambda_3 = 1$, $\lambda_4, \lambda_5, \lambda_6, \lambda_7 = 4$, and $\lambda_8 = 10$, (λ being the expected rate of occurrences, subscripts corresponding to X_1–X_8).
- Categorical: X_1–X_5 had 2 levels, probabilities equal to 0.2, 0.3, 0.4, 0.5, and 0.6, respectively. X_6 had 3 levels (probability of first and second category equal to 0.2). X_7 and X_8 had 7 levels (each with a probability of 1/7).
- Mixed: X_1 and X_2 normally distributed (with mean 60 and 120, respectively and a standard deviation of 15 each), X_3 Poisson distributed ($\lambda = 4$), X_4 exponentially distributed (rate parameter $\pi = 4$), X_5–X_7 binary (probabilities 0.3, 0.2, and 0.3, respectively), and X_8 categorical with 7 levels (each with probability of 1/7).

Generating the outcome variable with the regression model. Variable X_9 was considered the outcome variable and is a function of the predictors to which a normally distributed random error e is added. Let $\boldsymbol{D}$ be an n by m design matrix of variables X_1–X_8, with the *categorical variables transformed to dummy variables* (ie, $m > d$ for datasets with categorical variables). Let $e \sim \mathrm{N}(0, \sigma_e)$ and $X_9 = \boldsymbol{D}\boldsymbol{\beta} + e$, where the vector $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_m\}$ describes the effect of each variable in $\boldsymbol{D}$ on X_9.

All variables have a standardized effect of 0.3, meaning that 1 standard deviation unit difference on a variable in $\boldsymbol{D}$ has an effect of 0.3 on the outcome variable (for binary variables with probability p, the standard deviation is $p^*(1-p)$). An exception is the parameter β_1, the parameter of interest (previously referred to as β), for which the value was either a standardized effect of 0.3 or 0. Note that even when β equals 0, X_1 is correlated with X_9 through the other 7 variables. When the standardized effect of β was 0.3, the variance of e was such that R^2 is 0.3 and the same variance was used when $\beta = 0$. In all scenarios, the value of the intercept was zero.

Anonymization

MSPNs can be constructed with the `learn_mspn` function from the SPFLow Python package.[20] As input, we used **X**, an object indicating whether each variable was discrete or continuous, and the desired number of clusters (which was always set such that there was an average of 25 individuals per cluster). `learn_mspn` uses KMeans from sklearn.cluster to cluster the data with k-means++ and describes the data points in each resulting cluster separately for every variable with a histogram.[21]

To generate **X'**, the `sample_instances` function was used. This function first samples a cluster, with the sampling probability of each cluster corresponding to its weight, and next samples a value from each of the histograms connected to this cluster. We generate 100 000 records (as opposed to n) for the anonymized dataset, to ensure that (1) the information captured by the MSPN is accurately represented in the synthetic data and (2) information loss due to the sampling process is minimized. Standard errors are corrected for this difference, as will be detailed in the next subsection.

Several small modifications were made to `learn_mspn`. First, the Laplace smoothing of histograms was removed for a more accurate representation of the data. Second, we created clusters directly on the data instead of the standard procedure of recursive 2-fold clustering, which showed to lead to more accurate regression estimates. Third, we standardized the data before clustering. We chose the settings of `learn_mspn` (minimal number of instances and threshold for the product nodes) such that the algorithm always created a network with the structure described in the first subsection.

Quantifying information loss

To quantify the potential loss of information in the synthetic data generated from the MSPNs compared to the original data, a regression analysis was performed on each dataset. In this analysis, the outcome variable (X_9) was regressed on variables X_1–X_8. We allowed for estimation of an intercept (even though the intercept is zero in the data-generating model). We determined the number of repetitions required to detect a 5% mean difference between $\widehat{\beta}$ and $\widehat{\beta}'$ with a paired t-test for every scenario. The number of required repetitions to detect this difference was computed with power of 80% and α set at 5%, using the empirical standard deviation of difference. These numbers were rounded upwards. The number of repetitions per scenario can be found in Table 1. Whenever the averages $\widehat{\beta}$ and $\widehat{\beta}'$ do not differ significantly they are interpreted as "indistinguishable". We report bias (estimated by taking the average distance from the true value of β), the empirical standard error (ESE, standard deviation of $\widehat{\beta}$ or $\widehat{\beta}'$), and the average estimated standard error yielded by the regression analyses ($\widehat{SE}$). When the sample size of $\mathbf{X}'$ (n') differed from n, the corrected standard error was computed with the estimated standard error from the anonymized dataset ($\widehat{SE}'$): $\frac{\widehat{SE}'\sqrt{n'}}{\sqrt{n}}$.

Privacy evaluation

To evaluate privacy, 2 privacy metrics were computed both for the original data and for the MSPN. We presume that if the MSPN has successfully removed the individual-level associations that link background information to sensitive values, any synthetic data generated by sampling from the MSPN cannot contain this information either.

Privacy is computed for every variable X_j by evaluating for each individual in the dataset whether their value on this variable can be uncovered by using the values on all other variables. We use s to denote a sensitive variable and s_i to denote the value of the *ith* individual (or record) on this sensitive variable (ie, X_{ji} if X_j is the sensitive variable). Let $\mathbf{b}$ be the n by 8 matrix of background variables $\{X_1, \ldots, X_9\} \backslash s$ (ie, all variables except s) and let $\mathbf{b}_i$ denote the ith row in this matrix, containing the values on the background variables for individual i. Additionally, let dom(s) denote the domain of s, that is, the values the variable takes on, and let $v \in \mathrm{dom}(s)$, for example, the value *male* in the variable *gender*.

For unordered variables, we use the *Proportion of Alternatives Considered* (*PoAC*) by Kroes et al.[22] We slightly reformulate their definition so that it is suitable for the MSPN as well as for the original data and choose a threshold q of 0. It is defined as the proportion of values in dom(s) that do not equal s_i *and* have a nonzero *probability* given the values $\mathbf{b}_i$:

$$PoAC_{i,s} = \frac{|\{v \in (dom(s)\ s_i)|\ P(s = v|\boldsymbol{b}_i) \neq 0\}|}{|dom(s)|}. \quad (1)$$

Intuitively, we look at the distribution of the sensitive variable, which can take on a total of |dom(s)| distinct values, and we restrict this distribution such that the values on all *other* variables must correspond to $\mathbf{b}_i$. By computing how many of the values in dom(s) are still considered when conditioned on $\mathbf{b}_i$, we can quantify how effective this information is in uncovering s_i. The *PoAC* is a value between 0 (the worst case) and 1 (the maximum achievable). In the original dataset, the probability $P(s = v|b_i)$ is assessed by counting the number of individuals who have both the value v and the values $\mathbf{b}_i$. For the MPSN, we use the `log_likelihood` function in the SPFlow package to compute $P(s = v|b_i)$. Because this function never

Table 1. Bias, empirical standard error (ESE), and the average estimated standard error ($\widehat{SE}$) yielded by the regression analyses, for 4 different predictors distributions (normally distributed, count, categorical, and mixed discrete-continuous variables) for anonymized data (**X′**) with 100 000 records and various sample sizes (n) of original data (**X**)[a]

					X			**X′**		
Predictors	n	β	No. repetitions	Indistinguishable[b]	Bias	ESE β	$\widehat{SE}$ β	Bias	ESE β	$\widehat{SE'}$ β
Normal	1000	0	5500	✓	0	0.082	0.082	0.009	0.099	0.082
		0.3	5500	✓	0	0.082	0.082	0.001	0.099	0.082
	10 000	0	100	✓	0.004	0.026	0.026	0.004	0.029	0.026
		0.3	100	✓	0.004	0.026	0.026	0.003	0.030	0.026
	100 000	0	75	✓	0	0.008	0.008	−0.004	0.009	0.008
		0.3	75	✓	0	0.008	0.008	0	0.012	0.008
Count	1000	0	2500		0.001	0.077	0.078	0.020	0.087	0.078
		0.3	2500	✓	0.001	0.077	0.078	0	0.086	0.078
	10 000	0	150	✓	0.006	0.026	0.025	0.007	0.077	0.025
		0.3	150	✓	0.006	0.026	0.025	0.006	0.077	0.025
	100 000	0	100	✓	0	0.007	0.008	−0.001	0.011	0.008
		0.3	100	✓	0	0.007	0.008	0.001	0.010	0.008
Categorical	1000	0	150	✓	0.012	0.083	0.091	0.024	0.087	0.090
		0.75	150		0.012	0.083	0.091	0.026	0.086	0.090
	10 000	0	35	✓	−0.008	0.025	0.029	−0.007	0.026	0.028
		0.75	35	✓	−0.008	0.025	0.029	−0.007	0.025	0.028
	100 000	0	150	✓	0	0.007	0.009	0	0.012	0.009
		0.75	150	✓	0	0.007	0.009	0	0.010	0.009
Mixed	1000	0	4000		0	0.005	0.005	0.002	0.006	0.005
		0.02	4000		0	0.005	0.005	−0.004	0.005	0.005
	10 000	0	150	✓	0	0.002	0.002	0	0.002	0.002
		0.02	150	✓	0	0.002	0.002	0	0.002	0.002
	100 000	0	150		0	0.001	0	0.002	0.001	0
		0.02	150		0	0.001	0	0.002	0.001	0

[a]For the original data, the empirical and estimated standard errors are not always equal and the bias estimate can be nonzero. This is because for some scenarios, only a few repetitions were required to detect a 1% difference and because some distributions are skewed.

[b]Whether the P-value for a t-test comparing the empirical sampling distribution of the original and anonymized data is larger than or equal to 0.05, powered to find a 5% difference.

yields a probability of zero, we defined $P(s = v|b_i)$ to be nonzero if it was larger than or equal to $0.01*P(s_i|b_i)$.

For variables with a natural ordering, including continuous variables, we not only consider the number of sensitive values considered given certain background information but also how close these values are to the true sensitive value. Inspired by Li et al,[23] we compute the *expected deviation* (*ED*), which can be interpreted as the average distance from the true sensitive value, among values in the domain of s that have a nonzero probability given the background information:

$$ED_{i,s} = \sqrt{E[(F(s|\boldsymbol{b}_i) - s_i)^2]}, \quad (2)$$

where $F(x)$ is the estimated probability distribution of the sensitive variable. For the original data, we consider the values of s for all records that have background information equal to $\boldsymbol{b}_i$, compute the difference with s_i, and take the standard deviation of these differences. For the MSPN, the conditional distribution $F(s|\boldsymbol{b}_i)$ in the network is approximated by drawing 500 samples from the variable s, conditioned on $\boldsymbol{b}_i$ using `sample_instances`.

Note that in Equations (1 and 2), the values s_i and $\boldsymbol{b}_i$ and the subscript i always refer to the "true" values, that is, the values in the *original data*. Performing the required calculations involves a lot of computational power, particularly for larger datasets. Therefore, privacy was evaluated for 50 datasets per scenario for $n = 1000$ and 10 000 and for 10 datasets for $n = 100\,000$. For each dataset, privacy was computed for all variables for 1000 individuals. We consider either the original data or the MSPN to be "anonymized" when the appropriate privacy metrics are nonzero for all 1000 individuals for every variable. This means that it is not possible to infer the exact sensitive value of individual, if an attacker only has the information $\boldsymbol{b}_i$. The privacy values were averaged over the 1000 individuals and 10 or 50 repetitions to obtain 9 privacy values for both the original and the MSPN per simulation scenario.

RESULTS

Information loss

In Figure 3, the distribution of estimated regression parameters for the original and synthetic datasets are plotted. These plots show that these distributions are highly similar. In fact, the average estimated regression parameters with synthetic data were indistinguishable from the averages obtained from the original data, for 18 out of 24 scenarios. This can be seen from Table 1, where bias, standard errors and P-values are depicted. Performance was particularly good for $n = 10\,000$ and $n = 100\,000$. Even when the means of the distributions of regression parameters differed significantly between synthetic and original data, these differences were considerably smaller than the standard error (ie, the differences were small compared to

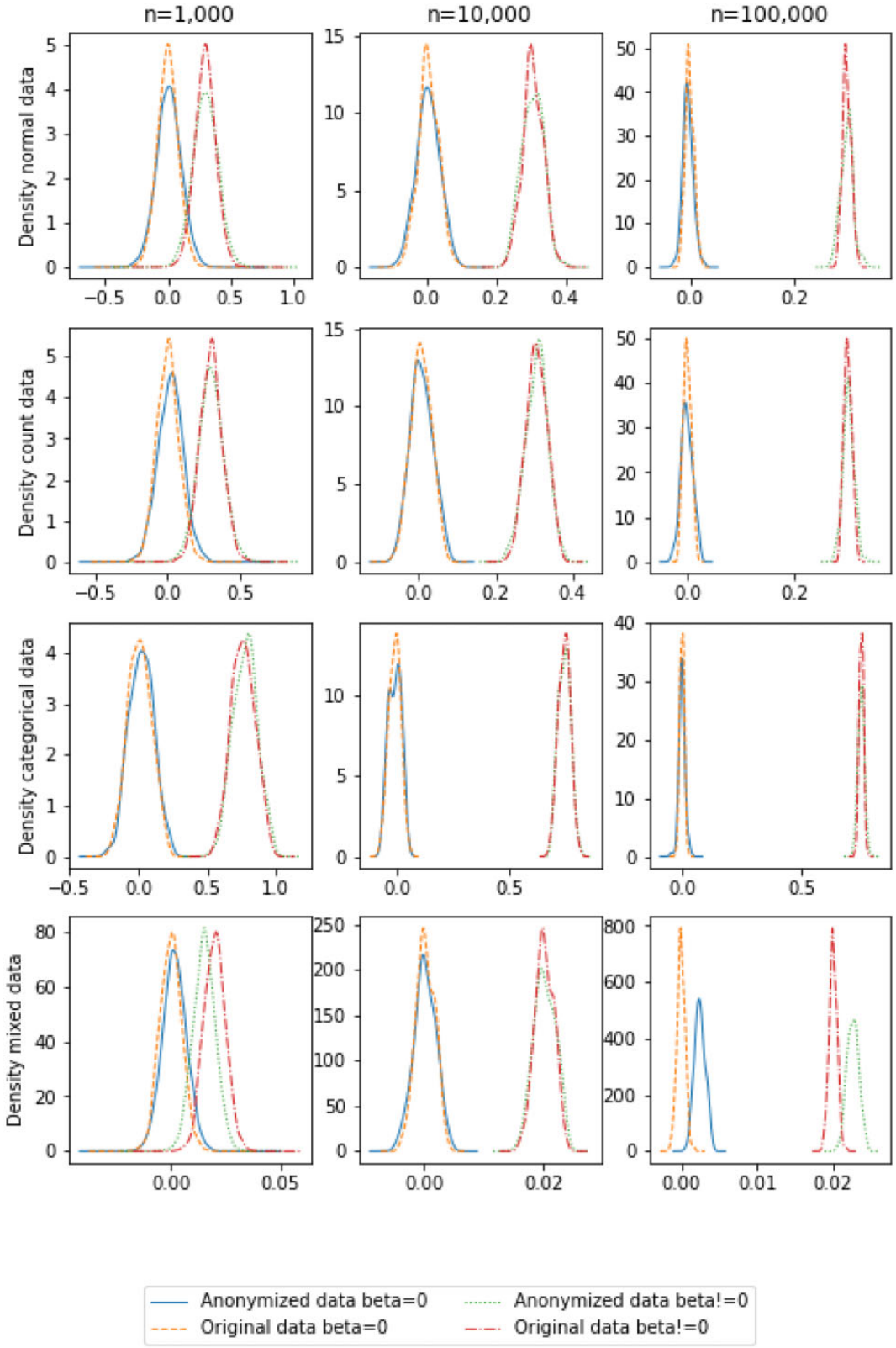

Figure 3. Simulation results of regression analyses with a continuous outcome. Distribution of estimated regression parameters is depicted for analyses with the original and anonymized data, for 4 predictor distributions (rows) and 3 sample sizes (columns) for 2 values of the parameter.

the width of the distribution with original data). An exception is scenarios with mixed predictors, where percentual differences were larger, but the absolute bias was very small (at most 0.004). ESEs were enlarged for the synthetic data compared to the original data in most scenarios, whereas the corrected estimated standard errors corresponded to those of the original data.

Table 2. Average privacy values and percentage of individuals considered anonymous for the original data (**X**) of sample size 1000 and corresponding MSPN when X_1 has an effect on X_9

Predictors			Variable								
			X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Continuous	SD		1	1	1	1	1	1	1	1	5.21
	ED	**X**	0	0	0	0	0	0	0	0	0
		MSPN	0.99	0.98	0.99	0.99	0.99	1.0	1.0	1.0	2.62
	% Anonymous	**X**	0	0	0	0	0	0	0	0	0
		MSPN	100	100	100	100	100	100	100	100	100
Count	SD		1	1	1	2	2	2	3.16	3.16	2.66
	ED	**X**	0	0	0	0	0	0	0.01	0.01	0.03
		MSPN	0.96	0.96	0.96	1.98	1.98	2.00	3.17	3.17	2.54
	% Anonymous	**X**	0	0	0	0	0	0	0	0	0
		MSPN	100	100	100	100	100	100	100	100	100
Categorical	No. levels[a] or SD		1*	1*	1*	1*	1*	2*	6*	6*	1.88
	PoAC[a] or *ED*	**X**	0*	0*	0*	0*	0*	0*	0*	0*	0.46
		MSPN	0.85*	0.90*	0.92*	0.94*	0.93*	0.91*	0.92*	0.92*	1.39
	% Anonymous	**X**	0	0	0	0	0	0	1	1	44
		MSPN	100	100	100	100	100	100	100	100	100
Mixed	No. levels[a] or SD		15	4	14	2.5	1*	2*	6*	1*	2.32
	PoAC[a] or *ED*	**X**	0	0	0	0	0*	0*	0*	0*	0
		MSPN	11.24	2.40	8.91	0.43	0.53*	0.50*	0.52*	0.51*	1.85
	% Anonymous	**X**	0	0	0	0	0	0	0	0	0
		MSPN	100	100	100	100	100	100	100	100	100

Note: Privacy results are depicted per variable, along with variable characteristics (standard deviation and number of levels). Higher values imply a higher level of privacy.

PoAC, Proportion of Alternatives Considered; ED, expected deviation.

[a]For unordered variables, the *PoAC* and number of levels are reported (as opposed to *ED* and SD), indicated with an asterisk.

Privacy

The average privacy values for original data and corresponding MSPNs are presented in Tables 2–4, for sample sizes of 1000, 10 000, and 100 000, respectively. We only present the results for the scenarios where β is unequal to zero, as the results are very similar for $\beta = 0$ (these can be found in Supplementary Appendix SA).

For the original data, sensitive values could be extracted with background information in all datasets. Particularly for data with continuous and mixed predictors, privacy could be breached for every individual, regardless of the size of the dataset. For categorical data, the privacy values increased considerably with the sample size, but for almost all variables a significant proportion of individuals remained at risk.

For the MSPNs, nearly all *PoAC* and *ED* values were nonzero, implying that the MPSN can indeed be considered a private representation of the data. An exception is the count datasets, where the *ED* was zero in very rare cases. Though this is not always visible in Tables 2–4, all % anonymous for count data are at least slightly below 100 (eg, 99.998%). This means that *if* the combination of background information (ie, 8 count values in the original data) were to be sampled by the MSPN for this small percentage of individuals, the sampled value for the ninth variable would always be equal to the true sensitive value.

The average *ED* was close to the original standard deviation for continuous and count data, indicating that having information on all 8 background variables does not significantly facilitate the deduction of the value on the ninth variable. Though *PoAC* and *ED* values were lower for mixed data, the MSPN meets our requirements to be considered "anonymized" for all normal, categorical and mixed scenarios (and the synthetic data by extension) and protects almost all tested individuals in count-valued data sets. (According to the definition in the section "Privacy evaluation".)

DISCUSSION

We proposed and evaluated an anonymization approach that combines the advantages of randomization, disassociation, and generalization. We generated synthetic data by sampling from an MSPN; a nonparametric probabilistic model of mixed discrete-continuous data that served as a private representation of the original data. Performance of the approach was evaluated with a simulation study, which showed that regression analysis with the synthetic data generated with MSPNs yielded results highly similar to those obtained from the original data. Furthermore, with respect to privacy, the synthetic data could be considered anonymized for at least 99% of tested individuals with the defined privacy threshold.

MSPNs have not yet been investigated in an approach to generate synthetic data. Additionally, for related approaches, the privacy-yielding factor is the generation of synthetic data, whereas the MSPN approach already protects privacy before the data are generated.[1–7] Furthermore, compared to most related work, the approach requires little pre-processing and no specification of parametric distributions by the data controller, while being suitable for mixed discrete-continuous data.

Unsurprisingly, the synthetic data resulted in a wider sampling distribution of regression parameters, because the MPSN is not an exact representation of the original data and the synthetic data are not reflective of every characteristic of the MSPN. Therefore, a correction needs to be applied to the estimated standard errors, in order to avoid increased Type I and Type II error rates. Though the aver-

Table 3. Average privacy values and percentage of individuals considered anonymous for the original data (**X**) of sample size 10 000 and corresponding MSPN when X_1 has an effect on X_9

Predictors			Variable								
			X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Continuous	SD		1	1	1	1	1	1	1	1	5.21
	ED	**X**	0	0	0	0	0	0	0	0	0
		MSPN	0.93	0.93	0.93	0.93	0.93	0.94	0.93	0.93	2.46
	% Anonymous	**X**	0	0	0	0	0	0	0	0	0
		MSPN	100	100	100	100	100	100	100	100	100
Count	SD		1	1	1	2	2	2	3.16	3.16	2.66
	ED	**X**	0	0	0	0	0	0	0	0	0.04
		MSPN	0.93	0.93	0.93	1.93	1.95	1.95	3.11	3.11	2.49
	% Anonymous	**X**	0	0	0	0	0	0	0	0	2
		MSPN	100	100	100	100	100	100	100	100	100
Categorical	No. levels[a] or SD		1*	1*	1*	1*	1*	2*	6*	6*	1.88
	PoAC[a] or *ED*	**X**	0.01*	0.01*	0.02*	0.02*	0.03*	0.01*	0.02*	0.02*	1.09
		MSPN	0.87*	0.91*	0.92*	0.92*	0.92*	0.91*	0.91*	0.91*	1.37
	% Anonymous	**X**	1	1	2	2	3	2	11	12	89
		MSPN	100	100	100	100	100	100	100	100	100
Mixed	No. levels[a] or SD		15	4	14	2.5	1*	2*	6*	1*	2.32
	PoAC[a] or *ED*	**X**	0	0	0	0	0*	0*	0*	0*	0
		MSPN	10.41	2.23	0.89	0.35	0.52*	0.37*	0.37*	0.51*	1.68
	% Anonymous	**X**	0	0	0	0	0	0	0	0	0
		MSPN	100	100	100	100	100	100	100	100	100

Note: Privacy results are depicted per variable, along with variable characteristics (standard deviation and number of levels). Higher values imply a higher level of privacy.

PoAC, Proportion of Alternatives Considered; ED, expected deviation.

[a]For unordered variables, the *PoAC* and number of levels are reported (as opposed to *ED* and SD), indicated with an asterisk.

Table 4. Average privacy values and percentage of individuals considered anonymous for the original data (**X**) of sample size 100 000 and corresponding MSPN when X_1 has an effect on X_9

Predictors			Variable								
			X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Continuous	SD		1	1	1	1	1	1	1	1	5.21
	ED	**X**	0	0	0	0	0	0	0	0	0
		MSPN	1.06	1.06	1.07	1.05	1.05	1.06	1.07	1.06	2.81
	% Anonymous	**X**	0	0	0	0	0	0	0	0	0
		MSPN	100	100	100	100	100	100	100	100	100
Count	SD		1	1	1	2	2	2	3.16	3.16	2.66
	ED	**X**	0	0	0	0	0	0	0.01	0.01	0.24
		MSPN	0.93	0.93	0.93	1.93	1.95	1.95	3.11	3.11	2.49
	% Anonymous	**X**	0	0	0	0	0.1	0	0.3	0.2	14
		MSPN	99	99	100	100	100	100	100	100	100
Categorical	No. levels[a] or SD		1*	1*	1*	1*	1*	2*	6*	6*	1.88
	PoAC[a] or *ED*	**X**	0.07*	0.08*	0.13*	0.12*	0.15*	0.09*	0.16*	0.15*	1.40
		MSPN	0.89*	0.92*	0.93*	0.93*	0.93*	0.91*	0.91*	0.91*	1.43
	% Anonymous	**X**	7	8	13	12	15	18	49	48	100
		MSPN	100	100	100	100	100	100	100	100	100
Mixed	No. levels[a] or SD		15	4	14	2.5	1*	2*	6*	1*	2.32
	PoAC[a] or *ED*	**X**	0	0	0	0	0*	0*	0*	0*	0
		MSPN	10.56	2.32	4.26	0.30	0.53*	0.53*	0.22*	0.52*	1.70
	% Anonymous	**X**	0	0	0	0	0	0	0	0	0
		MSPN	100	100	100	100	100	100	100	100	100

Privacy results are depicted per variable, along with variable characteristics (standard deviation and number of levels). Higher values imply a higher level of privacy.

PoAC, Proportion of Alternatives Considered; ED, expected deviation.

[a]For unordered variables, the *PoAC* and number of levels are reported (as opposed to *ED* and SD), indicated with an asterisk.

age regression parameters were indistinguishable for many datasets, for mixed predictors, the relative differences in regression parameters were larger. Overall, we conclude that the synthetic datasets were a viable substitute for the original data and that variable dependencies were highly similar.

To further increase the quality of the synthetic data, the clustering approach used to construct the MPSN could be improved. We created clusters using k-means on standardized data, but this is particularly suited for continuous data. Standardization also has a different effect on normal, skewed, and discrete data. Additionally, although minimizing within-cluster variance with k-means generally resulted in an adequate representation of the associations between the variables, preliminary analyses have indicated that more optimal partitions exist with respect to our anonymization objective.

Several strengths and limitations with respect to the experiments should be noted. First, interpretable privacy measures were used that can quantify whether sensitive information can be extracted with background information. This links in well with the concept of *identifiability*, which is critical in the European legislation regarding privacy (the General Data Protection Regulation). Additionally, because they can be used to measure privacy of both the original data and the MSPN, this allowed for a direct quantification of the privacy increase per individual resulting from the anonymization technique applied. A limitation of this approach, however, is its considerable computational burden. As a result, privacy was only computed for a subset of individuals. Although the results for these subsets consistently demonstrated that the MSPN removes opportunities for privacy attacks, a more efficient privacy computation algorithm is needed for a more complete analysis.

Another strength is that we systematically evaluated whether the desired properties of a frequently used and versatile analysis method (regression) are retained when substituting an original dataset with an anonymized dataset. We simulated different types of datasets and effect sizes. Furthermore, the MSPN had to grasp high-order dependencies, including confounding effects. This indicates that the MSPN can not only capture bivariate correlations, but also real-world complexities that are to be expected in medical data.

Further research is required to establish how and when the standard errors should be corrected and when the data are expected to yield unbiased results. Although the approach performs well for a variety of scenarios, future research can uncover how these results generalize to other types of analyses or different datasets (including "real" data), which were not studied in the experiments. Another research direction is how the number of clusters induces a privacy-utility trade-off, and which number of clusters is suitable, depending on the intended statistical analyses and desired levels of privacy. This trade-off is expected to be less favorable in datasets with a higher number of variables, few samples, or strong correlations, which may require a higher number of clusters to adequately capture the data structure and its joint distribution. At which point the user will have to compromise privacy or utility needs further research. Additionally, for time-series data, the approach needs to be extended to grasp within-subjects trends.

CONCLUSION

We presented an approach to create synthetic mixed discrete-continuous data, by sampling from constrained MSPNs. The proposed approach requires little user specification, removes most opportunities to extract sensitive information, and yields data that can reproduce regression analysis results. We conclude that the proposed method forms a promising new approach to generate private data.

FUNDING

This project is funded by the Sanquin Blood Supply Foundation (PPOC-16-27).

AUTHOR CONTRIBUTIONS

SKSK formulated the research proposal, designed and performed the simulation study and did most of the writing. MvL, RHHG, and MPJ supervised SKSK during each of these tasks and reviewed the drafts. All authors reviewed and approved the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We would like to thank A. Molina for helpful discussions and recommendations.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

All the data used for the simulations can be generated with the code in the repository https://github.com/ShannonKroes/MSPN_privacy.

REFERENCES

1. Torfi A, Fox EA. CorGAN: correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records. In: *The Thirty-Third International Flairs Conference*. 2020; Miami, FL, USA.
2. Piacentino E, Angulo C. Generating fake data using GANs for anonymizing healthcare data. In: *International Work-Conference on Bioinformatics and Biomedical Engineering*. Granada: Springer; 2020: 406–417.
3. Baowaly MK, Lin C-C, Liu C-L, Chen K-T. Synthesizing electronic health records using improved generative adversarial networks. *J Am Med Inform Assoc* 2019; 26 (3): 228–41.
4. Park Y, Ghosh J. PeGS: perturbed gibbs samplers that generate privacy-compliant synthetic data. *Trans. Data Priv* 2014; 7 (3): 253–82.
5. Drechsler J. Using support vector machines for generating synthetic datasets. In: *International Conference on Privacy in Statistical Databases*. Corfu: Springer; 2010: 148–161.
6. Kaur D, Sobiesk M, Patil S, *et al.* Application of Bayesian networks to generate synthetic health data. *J Am Med Inform Assoc* 2021; 28 (4): 801–11.
7. Li H, Xiong L, Zhang L, Jiang X. DPSynthesizer: differentially private data synthesizer for privacy preserving data sharing. In: *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*. Hangzhou: PubMed Central; 2014: 1677–1680.
8. Dwork C. Differential privacy: a survey of results. In: *International Conference on Theory and Applications of Models of Computation*. Berlin, Heidelberg: Springer; 2008: 1–19.
9. Lee J, Clifton C. How much is enough? choosing ε for differential privacy. In: *International Conference on Information Security*. Berlin, Heidelberg: Springer; 2011: 325–340.
10. Poon H, Domingos P. Sum-product networks: a new deep architecture. In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 2011: 689–690; Barcelona, Spain.

11. Sanchez-Cauce R, Paris I, Vegas FJ. Sum-product networks: a survey. *IEEE Trans Pattern Anal Mach Intell* 2021; 44 (7): 3821–39.
12. Molina A, Vergari A, Di Mauro N, Natarajan S, Esposito F, Kersting K. Mixed sum-product networks: a deep architecture for hybrid domains. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. New Orleans: AAAI Press; 2018: 3828–3835.
13. Puri V, Sachdeva S, Kaur P. Privacy preserving publication of relational and transaction data: survey on the anonymization of patient data. *Comput Sci Rev* 2019; 32: 45–61.
14. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. *l*-diversity: privacy beyond k-anonymity. *ACM Trans Knowl Discov Data* 2007; 1 (1): 3–es.
15. Sweeney L. k-anonymity: a model for protecting privacy. *Int J Unc Fuzz Knowl Based Syst* 2002; 10 (05): 557–70.
16. Xiao X, Tao Y. Anatomy: simple and effective privacy preservation. In: *Proceedings of the 32nd International Conference on Very Large Data Bases*. 2006: 139–150; Seoul, Korea.
17. Li T, Li N, Zhang J, Molloy I. Slicing: a new approach for privacy preserving data publishing. *IEEE Trans Knowl Data Eng* 2012; 24 (3): 561–74.
18. Terrovitis M, Liagouris J, Mamoulis N, Skiadopoulos S. Privacy preservation by disassociation. 2020. doi: 10.48550/arXiv.1207.0135.
19. Fung BC, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey of recent developments. *ACM Comput Surv* 2010; 42 (4): 1–53.
20. Molina A, Vergari A, Stelzner K, *et al*. SPFlow: an easy and extensible library for deep probabilistic learning using sum-product networks. 2019. doi: 10.48550/arXiv.1901.03704.
21. Arthur D, Vassilvitskii S. *k-means++: The Advantages of Careful Seeding*. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007: 1027–35; New Orleans, LA, USA.
22. Kroes SKS, Janssen MP, Groenwold RHH, van Leeuwen M. Evaluating privacy of individuals in medical data. *Health Inform J* 2021; 27 (2): 1–16.
23. Li J, Tao Y, Xiao X. Preservation of proximity privacy in publishing numerical sensitive data. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. Vancouver: Natural Sciences Publishing; 2008: 473–486.