An open generator of synthetic administrative healthcare databases

Thomas Guyet, Tristan Allard, Johanne Bakalara, Olivier Dameron

An open generator of synthetic administrative healthcare databases

Thomas Guyet[1], Tristan Allard[2], Johanne Bakalara[2,3], and Olivier Dameron[2]

[1] Institut-Agro/IRISA, Rennes

[2] Univ Rennes, Inria, CNRS, IRISA

[3] Univ Rennes, EA-7449 REPERES

Abstract : The recent development of data analysis provides opportunities for improving healthcare systems through analysis of health databases. However the thirst for data is conflicting with preserving the privacy of individuals. The generation of synthetic datasets may foster research on healthcare data analytics. It is mostly based on generative statistical models fitted on real data. Thus it still requires access to sensitive data. This article proposes a *probabilistic relational model* (Getoor *et al.*, 2007) fitted on publicly available datasets. Public healthcare statistics provide valuable information to mimic statistical distributions and do not hold sensitive personal data. More specifically, we propose to generate synthetic version of the national database of French insured patients. We do not only provide synthetic datasets, but a generator of datasets that can be used without any data access request. Experiments compare official statistics with those computed on synthetic datasets.

Mots-clés : Synthetic dataset Probabilistic Relational Model Privacy Epidemiology.

1 Introduction

The SNDS, formerly SNIIRAM, is a huge database (several Tb of data and about 700 tables) that contains information about healthcare reimbursements of about 60 million French insured patients (Tuppin & et al., 2017). This database is used to carry out epidemiological and medical-economic studies. Due to its sensitive medical content, identifying information (names, social security numbers) is removed or replaced by spurious information. Nevertheless, due to its richness, this database is not considered as anonymized. Consequently, its use is regulated and its access is restricted. Nonetheless, data owners would like to encourage research to derive the maximum possible benefit from it.

The access restriction to the SNDS hampers the possibilities of experimenting with new (machine learning) algorithms or studying their reproducibility. Without easy access to the data, researchers and engineers can not evaluate the effectiveness or the usefulness of their off-the-shelf algorithms. And, without hints about the potential utility of their algorithm, they are not spurred to request access to the database (which is a tedious process).

A solution is to generate and publicly release synthetic data (Jordon *et al.*, 2020; Tucker *et al.*, 2020). The generated data needs to mimic the real data in a structural way – in order to prepare and test preprocessing tools that access the raw database – and in a statistical way – in order to be confident that conclusions drawn on synthetic data will apply to the real data. Then, the challenge is threefold: 1) to generate relational data compliant with the original database schema, 2) to generate as realistic data distributions as possible and 3) to guarantee privacy preservation.

Synthetic data is a trending approach for affording widespread access to sensitive data (Raghunathan, 2021). A wide range of machine learning techniques are available to provide possible solutions. In particular, generative models (*e.g.* GAN (Yale *et al.*, 2019), Bayesian models (Chulyadyo & Leray, 2018; Liu, 2016; Tucker *et al.*, 2020), parameterized statistical models (Wu *et al.*, 2018)) can be trained to generate synthetic data. These techniques have been integrated in tools dedicated to synthetic data generation, such as Synthetic Data Vault (Xu *et al.*, 2019) or SynthPop (Nowok *et al.*, 2017) and are mainly applied on medical

datasets. Only a few approaches (Chulyadyo & Leray, 2018; Xu *et al.*, 2019) deal with relational data (*i.e.* multiple tables of a relational database). Modeling relation data raises more challenges for GAN models.

In the above approaches, the realism of the data is the primary objective, and they do not provide privacy guarantees. For instance, they may reveal sensitive information about real people with rare diseases (Stadler *et al.*, 2020). A trade-off must therefore be found between realism and privacy. GAN-DP (Qu *et al.*, 2019) or DL-DP (Abadi *et al.*, 2016) enhanced these approaches with differential privacy safeguards. But again, these approaches require primary access to real data.

Our approach also trains a statistical model to generate synthetic data but does not require access to the original database. The difference lies in the use of (aggregated) open data to fit the model. Our statistical model is based on a Probabilistic Relational Model (PRM) (Getoor *et al.*, 2007) which enables us to generate relational data compliant with the SNDS schema. Thus, the objective is not only to provide a synthetic dataset, but also to provide a generator that anyone can use (and adapt) with guarantees to reveal no more information than that available in open data.

2 Probabilistic relational models (PRM)

Probabilistic relational models (PRMs) extend Bayesian networks with the concepts of objects, their properties, and relations between them. It enables multiple instances of a class (*e.g.* multiple care events of a patient) to be modeled without having to ground the entire Bayesian network (one statistical variable for each event).

The following formal definitions are borrowed from Getoor *et al.* (2007). A *schema* for a relational model describes a set of classes (or *slots*), $\mathcal{X} = \{X_1, \ldots, X_n\}$. The set of descriptive attributes of a class X is denoted $A(X)$. Then, $X.A$ (resp. $x.A$) denotes the attribute A of the class X (resp. value of attribute A of an instance x), and $\mathcal{V}(X.A)$ is the domain of $X.A$. Each $X.A$ can be seen as a random variable. Similarly to Bayesian networks, a PRM represents conditional dependencies between variables using a set of parents $Pa(X.A)$. It distinguishes between dependencies for individual objects and dependencies on attributes of related objects (so-called *slot chain*). An *instance* of a relational model specifies for each class X, a set of objects by its attributes. A *relational skeleton* σ_r of a relational schema is a partial specification of an instance of the schema. More specifically, the skeleton of a PRM specifies the number of instances of each class. It may also fill object attributes with predefined values.

The joint distribution over the instantiations of a PRM, Π, for a relational skeleton, σ_r, is very similar to the chain rule for standard Bayesian networks.

$$\Pr(I|\sigma_r, \Pi) = \prod_{X \in \mathcal{X}} \prod_{A \in \mathcal{A}(\mathcal{X})} \prod_{x \in \sigma_r(X)} \Pr(x.A|Pa(x.A))$$

A synthetic database is a concrete sample of this joint distribution. Its generation has two goals: to generate a relational skeleton and to sample values to feed into the skeleton. The skeleton specifies a number of objects for each class. Chulyadyo & Leray (2018) consider two solutions: a relational skeleton with a nearly equal number of objects of each class, or the use of k-partite graph generation algorithm. Our solution is to add a random variable per class $X \in \mathcal{X}$, denoted $X.N$, that is the number of objects. These random variables are not descriptive attributes but are conditioned to variables of the PRM.

In the particular case of an acyclic conditional dependency graph, value sampling applies forward the probabilistic relational model from the roots of the covering tree of the dependencies ($\mathcal{R}$) to create an instance of the database. It leads to the sampling algorithm illustrated in Algorithm 1. Contrary to Chulyadyo & Leray (2018) who generate a skeleton beforehand, our skeleton is constructed during instance generation. For a slot, the number of its instances is drawn prior to generating instances of this slot.

Algorithm 1: Relational data sampling (acyclic parent graph)

$\mathcal{C} \leftarrow \mathcal{R}, \mathcal{A} \leftarrow \emptyset, \mathcal{D} \leftarrow \emptyset$
while $\mathcal{C} \neq \emptyset$ **do**
  foreach $X \in \mathcal{C}$ **do**
    $n \sim \Pr(X.N|Pa(X.N))$ `// draw a number of instances in` X
    repeat n **times**
      foreach *attribute* A *of* X **do**
        $x.A \sim \Pr(X.A|Pa(X.A))$
      add x to table X
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{X.A\}$
  $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{C}$
  $\mathcal{C} \leftarrow \{Y \in \mathcal{X} \setminus \mathcal{D} | \forall Y.A,\ Pa(Y.A) \subseteq \mathcal{A}\}$

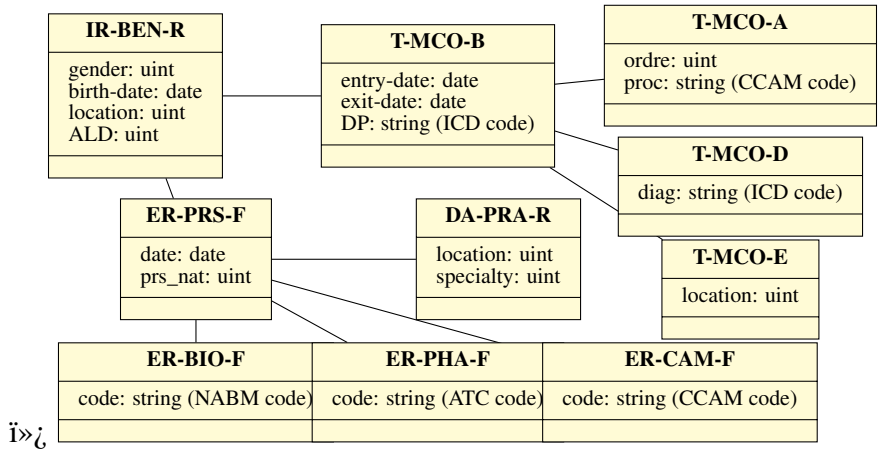

Figure 1: Part of the database schema whose content is synthetically generated. Tables are detailed in the text. NABM: French nomenclature of medical biology procedures, ATC: Anatomical Therapeutic Chemical drug classification system, CCAM: French common classification of medical procedures, ICD: International Classification of Diseases.

3 Probabilistic relational model of the SNDS

In this section, we first present the SNDS with a focus on information that is useful to conduct epidemiological studies. Then, we propose a PRM that specifies the generation of synthetic data.

The SNDS database has been set up for healthcare reimbursements. It has multiple facets: consultations (general practitioners and specialists), drug deliveries, biological procedures, hospitalizations, transport, dental care, nursing care. For each reimbursement, the database contains information about the patient, the providers (*e.g.* doctors, pharmacies) and the type of care through coding. It is worth noting that the SNDS is a reimbursement database. It does not contain exam results or medical reports.

The SNDS is composed of 692 tables, including 505 nomenclature tables which can be copied from the original database. Nomenclature tables contain descriptions of the standard codes (*e.g.* ATC codes, ICD codes), drug prices, care giver categories, etc. In addition, many administrative and economic tables are included, but not useful to generate for epidemiological purposes. Finally, based on the tables that are mostly queried by epidemiologists, we decided to discard specific types of care such dental care, optical care, physiotherapists, psychiatric stays and home care. Finally, 10 tables representing the core of a patient care pathway are actually generated. These tables are shown in Figure 1 and their meaningful attributes are detailed below. The names of tables correspond to the names used in the original schema.[1]

- `DA_PRA_R`: healthcare providers (doctors, pharmacies, nurses, etc.). A provider is

[1] `http://dico-snds.health-data-hub.fr/`

defined by their specialty and location (city code).

- `IR_BEN_R`: healthcare patients. This table contains information on the age, gender, city of residence, and information about the long-term illness reimbursement regime (so-called ALD). ALD provides information about chronic diseases (*e.g.* diabetes or cancers).
- `T_MCO_B`: Hospitalizations (short stay) for which there is an entry and a discharge date, as well as information about the main reason for hospitalization (ICD code).
- `T_MCO_D`/`T_MCO_A`: associated diagnoses/procedures. These tables contain respectively associated diagnoses and medical procedures for a hospitalization.
- `T_MCO_E`: hospitals and care institutions.
- `ER_PRS_F`: This table lists the care expenses (excluding inpatient care) which are reimbursed to a patient, after being prescribed by a doctor and provided by a care professional. Each reimbursement has a date and a code related to the nature of the care (*e.g.* drug delivery, consultation, biology). Biological procedures, drug deliveries and medical procedures are detailed in tables `ER_BIO_F`, `ER_CAM_F` and `ER_PHA_F`. Their main attribute is a standard code.

A PRM has been derived from the schema illustrated in Figure 1. It specifies the variables to sample and it models the conditional dependencies between them. The proposed model is illustrated in Figure 2. The overall structure of the model is close to the database schema. The patient class is at the center, and it is surrounded by slots describing care (medical procedures, drug deliveries, biological procedures and hospital stays). Note that slots for `T_MCO_E` and `T_MCO_D` tables are not represented in this schema for the sake of readability. Each slot holds a variable N which is the number of its record in the database, but it is not formally a slot attribute. It is worth noting that the choice of conditional dependencies has been guided by modeling purposes but it has been also constrained by data availability (see next section).

A patient is described by four random variables. Gender (G), residence location (L, a city code) and age (A) conditionally depend on the department (French administrative division) of residence D. This means that we model the population by a distribution $\Pr(G, A, L|D)$. The long-term illness reimbursement regime therefore depends on the three above variables. Let's now have a look at the ***Procedure*** slot which represents the medical procedures to generate in the `ER_CAM_F` table. C_{CCAM} is a code in the French classification of medical procedures. For instance, an imagery procedure will more likely be performed by a radiologist than by an oncologist. Then, the arrow from the triple G, A, L to C_{CCAM} states that we model the conditional probabilities $\Pr(C_{CCAM} \mid G, A, L)$ *i.e.* the probability of having a medical procedure of type C_{CCAM} during the year given the patient characteristics. The drugs delivered (C_{ATC}), reason for hospitalization (C_{ICD}) or type of biological procedures (C_{NABM}) depend on the patient only, but not on the care giver. For hospitalization, the in-hospital procedures (C_{HCCAM}) depend on the primary reason for hospital stay (C_{ICD}). It is the same for a related diagnosis (not represented). The ***Consult*** slot represents medical consultations (`ER_PRS_F`). The only attribute of this slot is the medical specialty (S) of the care giver that also depends on patient type (G, A, L). For instance, women are more likely to consult gynecologists than men.

The generation of synthetic data is configured by specifying a year of simulation and a list of departments whose population is to be mimicked. These parameters are used to generate a synthetic population and care professionals. Care professionals (physicians by specialty and care institutions) of these departments are extracted from public official lists.[2] As there is no cycle in the graph of conditional dependencies (see Figure 2), the Algorithm 1 is applied. It starts by generating a patient population. Then, for each patient, his/her possible ALDs are generated according to his/her age, gender and location. Then, treatments are generated as

[2]Official list available here: `http://open-data-assurance-maladie.ameli.fr/`

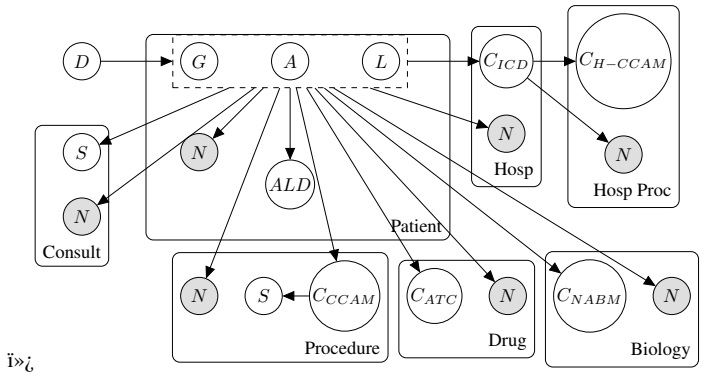

Figure 2: PRM corresponding to the SNDS schema (cf Figure 1). Each rounded-box corresponds to a slot. Each circle is a random variable. A circle in a box is an attribute of the slot except N (used for number of objects). The arrows show the parent relation of our model (conditional dependencies). The dashed box represents the joint variables A, G and L to avoid having three arrows each time.

follows: 1. draw the number of treatments of a given nature (hospitalization, visit, procedure, etc.) a patient had during the year using $\Pr(N|A, G, L)$; 2. draw the code (ICD, CCAM, ATC or NABM code) for each treatment using the probability of observing one code (*e.g.* $\Pr(C_{ATC}|A, G, L)$ or $\Pr(C_{CCAM}|A, G, L)$).

4 Conditional probability estimation from open data

To estimate the distributions of the conditional probabilities of our model, we propose to use data available in French open data repositories. The European Open Data Directive of 2019 requires administrations to make their data readily available, especially in the field of health. Most health datasets are composed of data aggregated from the real SNDS. Thus it enables the distribution of our PRM to be estimated: the *Open Damir* database[3] contains information about out-of-hospital reimbursements; the *Open Medic* database[4] describes expenditure by drug code, patient age, gender and location; and by prescriber specialty: the number of patients, the total amount of boxes delivered. The *Open CCAM* database[5] describes reimbursements of out-of-hospital medical procedures. The *Open Bio* database[6] describes reimbursements of medical biology procedures.

In addition, we use aggregated data[7] from hospitals about the primary, related and associated diagnoses and about the technical medical procedures carried out in hospitals. Finally, we use French population statistics[8] by city, gender and 5-year age group.

Challenges arise due to the incompleteness of the available information. Aggregate datasets are provided for analysis on three variables of the patient: age group, gender and location. Depending on the datasets, care deliveries are linked to characteristics of the provider, in particular his or her medical specialty. The aggregates are numbers of care events (*e.g.* number of drugs deliveries) or numbers of patients per care type given the value of these variables.

In practice, aggregates are rarely provided at a fine level of detail for all these variables, but more often for one or two of them. For instance, the total number of drug deliveries

[3] `https://www.data.gouv.fr/datasets/54de1e8fc751df388646738b`

[4] `https://www.data.gouv.fr/datasets/566e964188ee3875beaf0bf5`

[5] `https://www.scansante.fr/open-ccam/open-ccam-2019`

[6] `https://www.data.gouv.fr/datasets/58d3c14bc751df6883298f1c`

[7] Data platform of French healthcare institutions: `https://www.scansante.fr`

[8] French National Institute of Statistics: `https://www.insee.fr/en/statistiques`

(or biological procedures) is provided per department and per gender. Additionally, the total number of drug deliveries per department and per age group is available. This separation prevents possible re-identification of individuals. The joint distribution is reconstructed from the marginal distributions assuming conditional independence.

In the following, we present the principle of estimating distributions for the generation of the information in the $Procedure$ slot, *i.e.* the medical procedures which are performed by doctors including for example imaging procedures (radiology), surgical procedures, etc. We aim at estimating a quantity of procedures and their distribution:

- $\Pr(Proc.N|A, G, L)$: the probability of the number of procedures knowing the gender, age and department of residence of the beneficiary,
- $\Pr(C_{CCAM}|A, G, L)$: the probability of a CCAM code for an procedure knowing the gender, age and department of residence of the beneficiary.

Unfortunately, open data does not contain information to compute $\Pr(C_{CCAM}|\ A, G, L)$ exactly. On the one hand, the *OpenDamir* dataset gives the amount of procedures per group of procedures (radiology, ultrasound, other imaging, obstetrics, surgery, technical procedures, anesthesia) given the five-year age group, gender and location of the patient. This enables the $\Pr(G_A|A, G, L)$ to be computed where G_A is the group of procedures. On the other hand, we have information about the total amount of detailed CCAM procedures per group which gives $\Pr(C_{CCAM}|G_A)$. Then, we approximate the probability of a CCAM code as follows: $\Pr(C_{CCAM}|A, G, L) = \sum_{G_A} \Pr(C_{CCAM}|G_A) \times \Pr(G_A|A, G, L)$.

Furthermore, the distribution of the number of medical procedures ($\Pr(Proc.N|\ A, G, L)$) of a patient during the year is modeled by an exponential law parameterized by an average number of procedures given the age, gender and location of the patient. This average number of procedures is approximated using the same technique as above.

5 Experiments

All the tools presented in this article are available online[9]: 1) the procedures for loading open data (including recent SNDS schema and nomenclatures), 2) *Python notebooks* for preparation of the datasets and distribution estimations and 3) the simulation tool. These end-to-end tools enable the generation of a synthetic SNDS in a SQLite database. We have collected datasets for the year 2016. More data are available for this year than for other more recent years and 2016 is sufficiently recent to be representative of the current situation. The simulation is set up to generate ten populations of 10,000 people in 2019 and to mimic the patient population of the French Brittany region (four departments).

The objective of the experiments is to show that the generation of synthetic data reproduces the data distributions that are known from real open source data. For the sake of conciseness, we focus on population, drug deliveries and medical procedures. We present distributions of one synthetic dataset and the Jensen?Shannon divergence measure (JS) averaged over the 10 datasets. Chi-squared homogeneity tests assessed that there is no significant difference between the real and synthetic distributions. Extended results are available in the code repository on different database instances and for the other facets.

5.1 Population generation

Figure 3 compares the actual (blue) and synthetic (orange) distributions of the population by gender and age. For the real population, the numbers of people are proportionally reduced to the population size of the synthetic cohort. The synthetic and real data distributions are globally very close. For the age distribution, there is a tendency to underestimate the number of people aged over 70 but to overestimate those under 20. Nevertheless, the difference is small overall ($D_{KL} = 4.52 \times 10^{-3}$).

[9]Git repository: `https://gitlab.inria.fr/tguyet/medtrajectory_datagen`

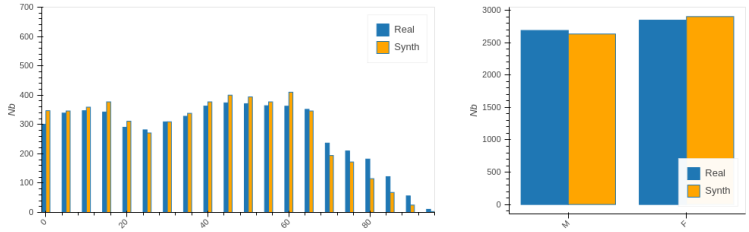

Figure 3: Distribution of the number of patients by age (left) and gender (right). The synthetic cohort is in orange and the real cohort is in blue.

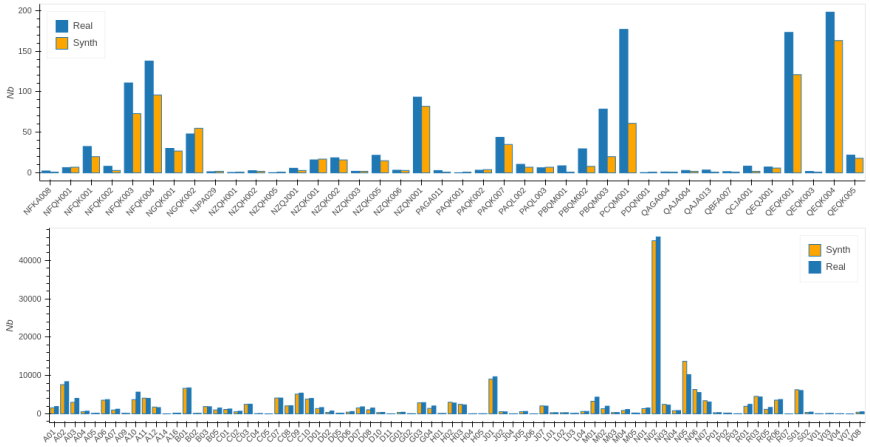

Figure 4: At the top: Distribution of CCAM procedures from NFKA008 to QEQK005 in synthetic population (in orange) and in the real national population (in blue). At the bottom: Distribution of the number of drugs per ATC group (level 2).

5.2 Distribution of drugs

Figure 4 (bottom) illustrates the results of the generation of drug deliveries. The numbers of deliveries are grouped by drug code from the second level of ATC. Therefore, the simulation enables both the numbers of deliveries and their distribution by type of drug to be rather faithfully represented (overall relative difference of 12% and $D_{KL} = 4.02 \times 10^{-3}$). Despite the wide disparity between the numbers of drugs per ATC code, the most frequently delivered drugs are also the most frequently delivered drugs in the synthetic cohort. These good results can be explained by the accuracy of the drug dataset, which details quantities by age, gender and region. The conditional distribution has been estimated without any additional assumptions.

5.3 Out-of-hospital procedures

Figure 4 (top) illustrates the actual and synthetic distributions of medical technical procedures (CCAM codes) for the codes from NFKA008 to QEQK005. The proportions of procedures are rather faithful to reality, but the quantities are globally underestimated. Moreover, for certain rather frequent procedures (*e.g.* PCQM001), the quantities of procedures present in the synthetic data are quite significantly lower (overall relative difference of 44% and $D_{KL} = 4.56 \times 10^{-2}$). These differences are due to the use of groups of procedures to obtain distribution conditionally dependent on gender and age.

These different experiments show that the data generation process reproduces the characteristics of SNDS data according to different facets of the care pathway. As expected, the facets for which the data are finely described in the open datasets (*e.g.* drugs) are more faithfully reproduced than for the less well described facets (medical procedures). Nevertheless,

it shows that the information available in open data makes it possible to generate synthetic data with meaningful epidemiological features.

6 Conclusion and perspectives

We proposed a tool for the generation of a synthetic SNDS. This approach does not require access to the original dataset and allows anyone to generate their own database. The advantages are to generate data respecting the complex SNDS schema and to reproduce real global characteristics of care used in epidemiological studies. The databases generated benefit from the privacy preservation measures applied on open data. Thus, generated datasets can be used to technically evaluate tools before transferring them to real data. A first perspective would be to enrich the tool with additional care facets (*e.g.* nursing care, odontology). The second area for improvement is to generate more consistent individual pathways. A first direction would be to explore the use of expert constraints (does not require using new data). The second direction would be to use individual data to statistically model the pathways in order to mimic them. The constraints of preserving privacy would then have to be taken into account.

References

ABADI M., CHU A., GOODFELLOW I., MCMAHAN H. B., MIRONOV I., TALWAR K. & ZHANG L. (2016). Deep learning with differential privacy. In *Proceedings of the Conference on Computer and Communications Security (SIGSAC)*, p. 308–318.

CHULYADYO R. & LERAY P. (2018). Using probabilistic relational models to generate synthetic spatial or non-spatial databases. In *Proceedings on the International Conference on Research Challenges in Information Science (RCIS)*, p. 1–12.

GETOOR L., FRIEDMAN N., KOLLER D., PFEFFER A. & TASKAR B. (2007). Probabilistic relational models. *Introduction to statistical relational learning*, **8**.

JORDON J., WILSON A. & VAN DER SCHAAR M. (2020). Synthetic data: Opening the data floodgates to enable faster, more directed development of machine learning methods. *arXiv:2012.04580*.

LIU F. (2016). Model-based differentially private data synthesis. *arXiv:1606.08052*.

NOWOK B., RAAB G. M. & DIBBEN C. (2017). Providing bespoke synthetic data for the UK longitudinal studies and other sensitive data with the synthpop package for R. *Statistical Journal of the IAOS*, **33**, 785–796.

QU Y., YU S., ZHANG J., BINH H. T. T., GAO L. & ZHOU W. (2019). GAN-DP: Generative adversarial net driven differentially privacy-preserving big data publishing. In *Proceedings of the International Conference on Communications*, p. 1–6.

RAGHUNATHAN T. E. (2021). Synthetic data. *Annual Review of Statistics and Its Application*, **8**(1).

STADLER T., OPRISANU B. & TRONCOSO C. (2020). Synthetic data – a privacy mirage.

TUCKER A., WANG Z., ROTALINTI Y. & MYLES P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, **3**(1), 1–13.

TUPPIN P. & ET AL. (2017). Value of a national administrative database to guide public decisions: From the système national d'information interrÃ©gimes de l'assurance maladie (SNIIRAM) to the systéme national des données de santé (SNDS) in france. *Revue d'ÉpidÃ©miologie et de Santé Publique*, **65**, 149–167.

WU H., NING Y., CHAKRABORTY P., VREEKEN J., TATTI N. & RAMAKRISHNAN N. (2018). Generating realistic synthetic population datasets. *Transactions on Knowledge Discovery from Data*, **12**(4), 1–22.

XU L., SKOULARIDOU M., CUESTA-INFANTE A. & VEERAMACHANENI K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, p. 7335–7345.

YALE A., DASH S., DUTTA R., GUYON I., PAVAO A. & BENNETT K. P. (2019). Privacy preserving synthetic health data. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.