ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Synthetic Data

Trivellore E. Raghunathan

Department of Biostatistics, School of Public Health, and Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48109, USA;
email: teraghu@umich.edu

Annu. Rev. Stat. Appl. 2021. 8:129–40

First published as a Review in Advance on September 22, 2020

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040720-031848

Keywords

full synthesis, partial synthesis, differential privacy, statistical inference, multiple imputation, Bayesian inference

Abstract

Demand for access to data, especially data collected using public funds, is ever growing. At the same time, concerns about the disclosure of the identities of and sensitive information about the respondents providing the data are making the data collectors limit the access to data. Synthetic data sets, generated to emulate certain key information found in the actual data and provide the ability to draw valid statistical inferences, are an attractive framework to afford widespread access to data for analysis while mitigating privacy and confidentiality concerns. The goal of this article is to provide a review of various approaches for generating and analyzing synthetic data sets, inferential justification, limitations of the approaches, and directions for future research.

1. INTRODUCTION

Data are essential ingredients for an open society, and society is well served when data—especially data collected using public funds—are collected using proper procedures, widely disseminated, and analyzed from multiple perspectives, and the results are debated to develop concrete policy decisions. The current explosion of the coronavirus disease 2019 (COVID-19) pandemic is a stark reminder that a society needs such information for making projections, planning, and developing a system of welfare for the society, not only from the public health perspective but also from the economic perspective. Thus, the need for detailed data on individuals in a society cannot be underestimated. Even though the need for data is critical, we, as a society, should not lose sight of the importance of the privacy and confidentiality of an individual's information; it is crucial to protect them from potential harm from an intruder who may glean their identity and, hence, any sensitive information from the released data set.

The data are collected from a set of respondents (this term encompasses any subject, such as an individual or a business unit, whose information is contained in the data set, which may be composed of survey or nonsurvey data) by various statistical agencies or their contractors, under the strict rules of privacy and pledges of confidentiality. Thus, any data collection is a contract between a respondent and the data collector. For this article, we assume that information that a data agency is seeking from a respondent is private and an informed consent process is required to collect that information. We stipulate that even if some information is available in the public domain (for example, court records), it is collected only after informed consent, because this public information cannot be linked to any information provided by the respondent to the data collector that is not available in the public domain. Of course, we may perform analysis solely based on publicly available information. For example, in the analysis of establishments, we may use information available in the public domain without any privacy concerns.

We define the right to privacy as a right of any respondent to keep the information private and not to divulge it to the data collector unless he/she/it wants to do so. Confidentiality is a pledge given to the respondent by the data collector that the private information given to the data collector will be kept confidential and used for statistical purposes. This pledge is typically interpreted as the steps that will be taken by the data collector to make sure no one will know that specific information has been provided by a given respondent. Numerous laws have been passed to make this pledge legally enforceable, with stiff punishments for violation.

Under this rubric, we have become accustomed to various government statistical agencies collecting data through surveys or administrative sources (e.g., medical claims and credit card transactions) and making them available for statistical analysis, either in the form of public-use data sets or through the mechanism of data use agreements. Some confidential data are made available for statistical purposes through Research Data Centers with strict access protocols.

Statistical disclosure control (SDC) is the process by which (just as informed consent is the process to maintain the right to privacy) the confidentiality pledge is to be fulfilled—that is, the steps to be taken by the data collector to ensure that the risk of gleaning the identity of any respondent from the released data is minimal or negligible. Thus, maintaining the privacy of the respondents and the confidentiality of the information provided by them has also been a long-term concern. When data are made available as public-use data sets, agencies have taken a variety of SDC steps to protect the privacy and confidentiality of respondents.

Dalenius (1974, 1977, 1978) provides a comprehensive review of the issues and discusses various approaches that have been used for limiting disclosure. The National Academy of Sciences has issued several reports (NRC 1993, 2003, 2005; Natl. Acad. Sci. Eng. Med. 2017) on privacy, confidentiality, and data access; these topics have also spawned several special issues of *Statistica*

Neerlandica (Keller & Kooiman 1992) and the *Journal of Official Statistics* (Duncan 1993, Fienberg & Willenborg 1998), for example, and a myriad of articles addressing the issues of privacy, confidentiality, disclosure limitation, data access, and so on.

To be concrete, assume that the data collected from a set of respondents consist of three types of variables: I, personally identifying variables such as names, addresses, telephone numbers, or social security numbers; X, variables of interest like demographics or some geographical indicators; and Y, substantive variables of interest for a particular study or survey, such as disease status. The candidate data for statistical purposes consist of $D = (X, Y)$. The disclosure problem concerns some intruder correctly finding some values in I for any individual when D is released.

Most, if not all, methods involve altering the data, $D = (X, Y)$, to prevent any potential intruder from gleaning the identity of an individual respondent. They include swapping some values of a few variables, mostly in X, between one or more respondents or coarsening the variables in X (for example, instead of providing age in 1-year intervals, create age groups in 5- or 10-year intervals; instead of listing zip codes, provide a larger geographical area such as a county or state). Another strategy is to suppress some unique combination of variables when producing a table. These approaches are designed to hide sensitive subjects from any potential intruder. Some analyses of the altered data may yield inferences different from the original data, but they may be minimal if the alterations are not substantial [see Lane et al. (2001) for details about various procedures in vogue at various statistical agencies].

2. SYNTHETIC DATA

Much of modern technology relies on the process called synthesization, a chemical process that, by human agency, emulates certain properties of a naturally occurring material. Such technology allows one to widely use key ingredients or products while preserving natural resources. For instance, indigo plants, which are native to southeast Asia, would have become extinct if the natural dye originally developed by Levi Strauss were used to create all the jeans in the world. The idea of synthetic data sets is similar. A statistical process is used to extract information from an actual data set collected from a set of respondents and is reexpressed as a collection of artificial or synthetic data sets for public consumption. This allows wide dissemination of the informational content of the actual data set and, at the same time, limits the exposure to potential inadvertent or malicious disclosure of sensitive information about the respondents.

As indicated earlier, almost all SDC methods are in some way altering the data to prevent disclosing the identity of a respondent. Duncan & Pearson (1991) give a systematic conceptualization of altering or masking the data using a transformation. Let D be the actual data set, with n rows representing the respondents and p columns representing the variables. Define $D^* = ADB + C$, where A is an $n \times n$ matrix that transforms the cases, B is an $n \times p$ matrix that transforms the variables, and C is an $n \times p$ matrix that adds perturbations, or a blurring factor, to the entries of ADB. Cox (1994), Fienberg et al. (1998), Gouweleeuw et al. (1998), and Fuller (1993) give detailed discussions relating many of the SDC methods to the Duncan–Pearson formulation of masking the data set for disclosure control and its impact on the informational content.

Inference based on D^*, however, could be quite different than that based on D. Construction of inferences from D^* about the estimand, a function of the parameters in the probability distribution $\Pr(D|\theta)$, is not straightforward [see Little (1993) for a discussion from the likelihood-based-analysis perspective] and depends upon the information released about A, B, and C. Of course, the data producer may not want to share information about A, B, and C to prevent reengineering to find D and then using it to find I. Thus, these methods may reduce the risk of disclosure but may severely affect the utility of the masked data from the analyst's perspective.

The idea of creating synthetic data to obtain valid statistical inferences and achieve disclosure control was first proposed by Rubin (1993) as a discussion of Jabine (1993), who described the then-current SDC approaches used by the federal statistical agencies. In developing SDC methods, there is a trade-off between taking steps to prevent disclosure of the identity of the respondents and preserving the inferential utility of the data collected. We consider the three approaches for creating synthetic data sets (full, partial, and selective) that represent various levels of balancing between these two goals. Whichever approach one may use to create synthetic, partial, or full data sets, the fundamental goal is to maintain their usefulness. The main use for any data collected is statistical inference. A frequentist may construct point and interval estimates of a certain population quantity or a parameter in a model posited as a data generation mechanism. Similarly, a Bayesian may construct the posterior distribution of the same population quantity or the parameter. Thus, this notion of valid inferences from the synthetic data needs to be understood and ensured in any process that creates such synthetic data sets. Section 3 provides details about the notion of valid inferences and methods to assess the usefulness or appropriateness of using synthetic data for statistical inferences.

As before, suppose that D is an $n \times p$ data matrix on n subjects with p variables. The basic idea of fully synthetic data sets is to generate a collection, $D_{\text{syn}} = \{D_*^{(l)}, l = 1, 2, \ldots, M\}$, of M data sets that is based on the statistical informational content in D. This idea is rooted in the Bayesian principles of the analysis of finite population survey data and a multiple imputation approach for handling the missing values. Raghunathan et al. (2003) fully describe this approach, develop methods for the analysis of synthetic data, and evaluate the methodology using simulated and actual data sets. We discuss this approach in Section 4.

Little (1993) suggests partially masking or synthesizing data instead of a full synthesis. The partial synthesis can be of some rows in the data matrix (all the records of some subjects are synthesized to protect them) or some columns (some key variables for all the subjects are synthesized to protect them). The key variables may be demographic variables that can potentially identify the respondents or highly sensitive variables such as income. Reiter (2003) develops the methodology for analyzing partially synthetic data sets. Little et al. (2004) generalize this approach by selecting a subset of both rows and columns to be synthesized for the SDC (selective synthesis). We discuss partial synthesis in Section 5.

The digital revolution is more than 50 years old, but the recent advances in computing power to harness digitally collected information have ushered in an era of big data. Availability of data from social media, electronic administrative records, business transactions, and other private or public sources is constantly increasing the chances of disclosure of sensitive information about respondents. In response to this threat, a new line of research is emerging using the notion of differential privacy (Dwork et al. 2006). Under this approach, no data are released, but results from a user-specified sequence of analyses are made available. The results are altered to prevent reconstruction of the data set D from the ensemble of analyses. These differentially private analyses are aimed at protecting the data, regardless of what information may be available to an intruder. This approach is being modified to release differentially private synthetic data sets, which are more suitable for statistical practice. This approach, still in its infancy, is briefly described in Section 6.

We briefly discuss some available software packages for creating and analyzing synthetic data sets in Section 7. Finally, Section 8 of the article concludes with a discussion of potential pitfalls of SDC approaches and advice to potential users of these approaches for protecting the confidentiality promised to the respondents.

3. VALID STATISTICAL INFERENCES

Whatever SDC methods are used to alter the data, we would like to ensure that the inferences from the altered data are valid. From the frequentist perspective, the validity can be formulated in terms of the repeated sampling properties of the point and interval estimates. For example, suppose θ is a scalar parameter of interest, $\widehat{\theta}_{\text{syn}}$ is the estimate based on synthetic data, and the corresponding $100(1-\alpha)\%$ interval estimate is $(L_{\text{syn}}, U_{\text{syn}})$. At a minimum, we would expect $\widehat{\theta}_{\text{syn}}$ to be a consistent estimator of θ, and preferably unbiased for θ. Similarly, across repeated samples, we would require $\Pr(L_{\text{syn}} \leq \theta \leq U_{\text{syn}}) \geq 1-\alpha$, preferably as close to $(1-\alpha)$ as possible. That is, the inference procedures using the synthetic data sets are calibrated from the repeated sampling perspective.

We may also be interested to know how similar the inferences are using the synthetic and actual data sets. Suppose $\widehat{\theta}_{\text{act}}$ is the actual estimate, with the corresponding interval estimate as $(L_{\text{act}}, U_{\text{act}})$. The data producer can investigate, for a selected set of estimands, the differences $||\widehat{\theta}_{\text{syn}} - \widehat{\theta}_{\text{act}}||$. Similarly, one would desire that the interval estimates overlap to a considerable extent, except that synthetic data interval estimates may be slightly wider due to loss of efficiency. Such investigations will convey confidence in using the synthetic data for constructing inferences.

From the Bayesian point of view, suppose that $\pi(\theta|D_{\text{syn}})$ and $\pi(\theta|D)$ are the posterior densities of the parameter θ, based on synthetic and actual data sets, respectively. A desirable feature is to ensure that the features in $\pi(\theta|D)$ are also exhibited in $\pi(\theta|D_{\text{syn}})$. A useful scalar summary is the Kullback–Leibler distance,

$$\int \log \frac{\pi(\theta|D_{\text{syn}})}{\pi(\theta|D)} \pi(\theta|D) \mathrm{d}\theta.$$

Given that most Bayesian analyses are based on draws from the posterior distribution of the parameters, the draws from $\pi(\theta|D)$ and $\pi(\theta|D_{\text{syn}})$ may be compared across models to assess the validity of inferences from synthetic data sets.

The data producer will have to take steps, in some sense, to assure that $\Pr(D) \approx \Pr(D_{\text{syn}})$ for valid analyses and identify those instances where D_{syn} cannot be used to draw inferences. As an example, suppose that all the variables in D are categorical. To create synthetic data sets, suppose that a log-linear model, suppressing some higher-order terms, is fitted and M independent draws from the corresponding posterior predictive distribution or synthetic tables are released. The informational content in the released data is the factors included in the log-linear model, and therefore, D_{syn} cannot be used to infer about the factors not included in the model.

4. FULL SYNTHESIS

The full synthesis approach uses the basic principles of a Bayesian analysis of survey data and handling of missing data through multiple imputation (Rubin 1987). In a Bayesian analysis of sample survey data from a finite population, the nonsampled portion of the population is treated as missing data, and the Bayesian inference essentially involves constructing the posterior predictive distribution of the unobserved values in the population conditional on the observed values. Usually a model, which can be nonparametric, semiparametric, or fully parametric, is used to construct this posterior predictive distribution. Thus, Bayesian inference, in this context, treats nonsampled values as missing, and therefore, the multiple imputation approach can be used by actually drawing several sets of nonsampled values from this posterior predictive distribution that, when appended to the sample, create synthetic populations. The Bayesian inference can be constructed by analyzing the multiple synthetic populations. If the number of synthetic populations is large, the procedure just outlined is a Monte Carlo implementation of a Bayesian inference for survey data. When the number of imputations is small, the multiple imputation analysis of synthetic populations can be viewed as an approximation of the Monte Carlo implementation of a Bayesian inference. This is a useful inferential tool for an analyst using a particular model for a subset of variables. For large samples, such Bayesian inferences also have desirable repeated sampling properties.

Because the finite population is usually very large in practical problems, manipulating multiple synthetic populations is not practical. In the confidentiality context, Rubin (1993) proposes that a sample from each synthetic population be released instead. This step introduces an additional source of variability that needs to be incorporated in the analysis of multiple synthetic samples. Raghunathan et al. (2003) develop procedures for analyzing multiple synthetic samples, provide analytical results, demonstrate the validity of inferences from a repeated sampling perspective, and show that the inferences from the original data and full synthesis data sets are comparable for simple models such as multiple normal linear regression and tobit models. Reiter (2002) also demonstrates that this approach results in valid inferences for complex survey designs if the model used to construct predictive inference incorporates these complex design features.

Briefly, suppose that M synthetic samples are released as public-use data sets. The analyst is interested in the estimand θ (e.g., percentile, mean, or regression coefficient). Let $\widehat{\theta}_l, l = 1, 2, \ldots, M$ denote the point estimate and v_l be its variance estimate (i.e., square of the standard error). The synthetic estimate is

$$\widehat{\theta}_F = \frac{1}{M}\sum_l \widehat{\theta}_l, \qquad 1.$$

and the variance estimate is

$$T_F = (1 + M^{-1})b - \bar{v}, \qquad 2.$$

where $b = \sum_l (\widehat{\theta}_l - \widehat{\theta}_F)^2/(M-1)$ and $\bar{v} = \sum_l v_l/M$. The variance estimate in Equation 2 can be occasionally negative, especially if M is small. For example, when $M = 10$, the estimate is rarely negative. Raghunathan & Rubin (2000) suggest an ad hoc fix, $T_F = k\bar{v}/n$ whenever the estimates given in Equation 2 is negative, where k is the size of the synthetic sample. This ad hoc fix was evaluated by Reiter (2002), who concluded that the resulting confidence intervals are well calibrated.

5. PARTIAL SYNTHESIS

Arguably, creating synthetic populations is a complex task in a practical setting, as survey data may involve several hundred variables of different types. We may select a set of variables called keys, which could be used by an intruder to identify a respondent, to be synthesized instead of all the variables. That is, only the key variables change across synthetic samples, and the nonkey variables remain the same. Abowd & Woodcock (2001, 2004) use a similar idea, except that they split the data set into sensitive and nonsensitive variables and synthesized only sensitive variables.

Suppose that the variables in the survey data can be split into two groups: keys, X, and nonkeys, Y (sensitive and nonsensitive variables, in Abowd–Woodcock terminology), where $D = (X, Y)$ is the actual observed data. Partial synthetic samples are $D_{\text{psyn}} = (X_*^{(l)}, Y; l = 1, 2, \ldots, M)$, where M, as before, is the number of synthetic samples and $X_*^{(l)}, l = 1, 2, \ldots, M$ are independent draws from the posterior predictive distribution, $\Pr(X|D)$.

Reiter (2003) demonstrates that the variance estimate given in Equation 2 is not applicable while analyzing multiple partially synthetic samples and develops a combining rule for estimating the variance as

$$T_P = \bar{v} + b/M,$$

and the point estimate remains the same.

To further minimize the amount of data to be modified for confidentiality protection, Little et al. (2004) suggest that only key variables for sensitive cases and a subset of nonsensitive cases be multiply synthesized. That is, this approach selects not only the variables but also the cases to be synthesized. They call this procedure selective multiple imputation of keys, but it falls under the rubric of partial synthetic samples.

There are several examples of surveys creating and publishing synthetic data sets. Readers are directed to, for example, Kennickell (1999) for an application to the Survey of Consumer Finances, Abowd et al. (2009) for an application to the Longitudinal Employer- Household Dynamics (CMS 2013) regarding synthetic Medicare claim files, and Sakshaug & Raghunathan (2010) for an application to the American Community Survey. A potential use of the synthetic data strategy to expand research access to tax data is proposed by Burman et al. (2018). Another, rather ambitious, project is adopted by Walonoski et al. (2018), who use substantive knowledge, published literature, and some microdata to create fully synthetic patients and their electronic health records. The validity of inferences from such enterprises still needs to be investigated.

6. DIFFERENTIAL PRIVACY

As mentioned earlier, the amount of information about businesses or individuals that may be available to potential intruders is ever increasing. It is not clear whether the synthetic data methods discussed above are able to prevent disclosure, regardless of the information available to intruders. Thus, a better strategy is to build the prevention into our synthetic data generation process. The notion of differential privacy has brought a rigorous process of conceptualizing the risk and integrating the prevention into the dissemination process.

There are two directions in differential privacy-based dissemination. In the first approach, no data will be released; only differentially private analysis results will be made available based on user requests. This approach has a total privacy budget, and each user request spends a portion of this budget. Once the entire budget is spent, no further analysis will be made available. Though this approach may work for some statistical agencies that are charged with producing tables, means and proportions, or descriptors and summaries of the data, it is very unrealistic from the statistical practice perspective. Any analyst knows that a real statistical investigation involves doing exploratory data analysis, developing models, assessing model fit, refining the models, and, once they are finalized, using the models for constructing inferences such as point and interval estimates or the posterior distributions of the parameters of interest.

The second, and more promising, approach is to generate differentially private synthetic data sets. Several methods have been developed (see, e.g., Wasserman & Zhou 2010, Reiter & Dreschler 2010, Bowen & Snoke 2019). The goal of either approach is to ensure that reconstruction of the original D is not possible from the disseminated synthetic data sets but to afford the broader analytical goals of multiple users.

As before, suppose that D is the $n \times p$ matrix of p variables on n subjects and D_{-i} is the $(n-1) \times p$ matrix without the subject i. Let synthetic data mechanisms involve generating D^* from a distribution $\Pr(z|D)$. The mechanism is defined to be ϵ-differentially private if

$$\frac{\Pr(D^* \in A|D)}{\Pr(D^* \in A|D_{-i})} \leq e^{\epsilon}$$

for every $i = 1, 2, \ldots, n$. If $f(z|D)$ is the corresponding density function, then

$$\sup_z \frac{f(z|D)}{f(z|D_{-i})} \leq e^{\epsilon}.$$

That is, there is very little effect on the synthetic data mechanism, whether any one particular subject is included or not.

We consider a simple example to this type of synthetic data mechanism. Suppose that $p = 1$, and D_i is binary, taking the value 1 or 0. Let D^* be generated from $\Pr(z|D)$ where $f(z|D)$ is the corresponding density or mass function. Note that, though D is binary, D^* need not be. Consider an individual i, and let the odds of an intruder correctly identifying the value of D_i as 1, without knowing D^*, be $O_i = \Pr(D_i = 1)/\Pr(D_i = 0)$. Let the odds of identifying the value of D_i as 1, when D^* is released, be $O_i^* = O_i(D^*) = \Pr(D_i = 1|D^*)/\Pr(D_i = 0|D^*)$. Using Bayes' theorem and an approximation $\exp(\epsilon) \approx 1 + \epsilon$, any differentially private synthetic mechanism needs to satisfy $1 - \epsilon \leq O_i^*/O_i \leq 1 + \epsilon$ for all $i = 1, 2, \ldots, n$. McClure & Reiter (2012) provide an example for the binary data. Let $\alpha = \beta = [\exp(\epsilon/n) - 1]^{-1}$ and generate D_i^*, $i = 1, 2, \ldots, n$ as independent Bernoulli random variables with probability

$$g = \left(\sum_{j}^{n} D_j + \alpha\right) / (n + \alpha + \beta).$$

Another example is the exponential mechanism proposed by McSherry & Talwar (2007). Again, focusing on $p = 1$, let F_D be the distribution function of the actual data and F_{D^*} be the distribution function of the synthetic data mechanism. Suppose that $\Delta(D, D^*) = \sup_t |F_D(t) - F_{D^*}(t)|$. Let $\Delta_{\max} = \sup_i \Delta(D, D_{-i})$. Define

$$f(D^*|D) = \exp\left(-\frac{\epsilon\,\Delta(D, D^*)}{2\,\Delta_{\max}}\right).$$

Developing a general-purpose differentially private synthetic data mechanism with different kinds of variables is a complex task. Bowen & Snoke (2019) provide a review of parametric and nonparametric approaches for creating differentially private synthetic data sets. Most of these methods are geared toward tabular data. However, the evaluation metrics are not the properties of the inferences, like coverage properties of confidence intervals from the synthesized data, which is of utmost interest to substantive researchers.

In this era of big data, we have another challenge of processing multiple large synthetic data sets. When n is large, we may need to consider releasing $k \times p$ dimensional synthetic data sets where k is considerably smaller than n. DuMouchel et al. (1999) discuss an approach they call data squashing, which may be useful to reduce the burden of analyzing synthetic data sets. In particular, the likelihood-based approach discussed by Madigan et al. (2002) may be useful to integrate with a parametric approach for creating synthetic data sets. We also have situations where the number of variables p is very large compared with n. We may have to consider the opposite, an expansion where k is much larger than n, to facilitate routine analysis of synthetic data sets using available statistical software packages.

7. SOFTWARE PACKAGES

Abowd & Woodcock (2001, 2004) use the sequential regression approach, which has been proposed for handling missing data (Raghunathan et al. 2001), to generate synthetic data. Specifically, suppose that $D_1, D_2, \ldots, D_p$ are the p variables in the data set. The joint distribution, $\Pr(D_1, D_2, \ldots, D_p)$, is modeled as

$$\Pr(D_1)\Pr(D_2|D_1)\ldots\Pr(D_p|D_1, D_2, \ldots, D_{p-1}).$$

Each conditional distribution is then modeled as a regression model, and the synthetic data are generated as draws from the sequence of the corresponding posterior predictive distributions.

The conditional distributions may be parametric, semiparametric, or nonparametric. An attractive feature of this approach is that model development involves p standard regression analyses and diagnostics. Furthermore, the missing values in the original data set can be imputed during the synthetic data generation process, as described by Raghunathan et al. (2018a,b).

The sequential regression approach has been implemented in the R package `synthpop` (Nowok et al. 2016) and in Version 0.3 of the software package IVEware (Raghunathan et al. 2018a,b) (**www.iveware.org**). When the original data have missing values, both of these packages can be adapted to impute the missing values and create synthetic data sets simultaneously. In addition, IVEware can handle data from complex sample surveys.

As new methods are being developed to make the synthetic data mechanism differentially private, the corresponding software routines are made available in the GitHub repository (see, e.g., **https://github.com/topics/synthetic-dataset-generation**). Most of these packages use Python or R and may need considerable modification by the user. Surendra & Mohan (2017) provide a survey of methods for creating privacy-preserving partial or full synthetic data sets, using either real or simulated data sets.

8. LIMITATIONS AND DISCUSSION OF FURTHER RESEARCH

There is no doubt that making data accessible to the research community for analysis will lead to sound policy decisions and benefit society. At the same time, it is important to maintain the privacy and confidentiality of respondents who are trusting the data collector with their information. It is also clear that the potential use of available external information to either intentionally or unintentionally identify and disclose respondents is ever increasing. Thus, the methods for making data accessible need to be evolving and continually refined in response to threats. The synthetic data framework is useful in that regard. In fact, all statistical disclosure methods, which have a long history, involve altering the data and may be construed as synthesization.

It is important that any synthetic data development does not lose site of the ultimate goal of drawing valid statistical inferences. If this goal is not maintained, then there is no purpose for collecting the data. Of course, the data can be used for internal secretive or confidential analysis, and the results used for developing policies, but that is not a prescription suitable for an open democratic society. Releasing tables or analyses that are so altered that they deviate enormously from the actual data results is also an useless endeavor.

For the synthetic data approach, full or partial, to succeed in practice, the synthetic populations should be created to give valid inferences for a wide variety of models. That is, the same set of draws created by the data collector should provide valid inferences for a wide collection of analyst models. Obviously, the simulations created under a model will reflect its properties. In the missing data literature, it is well known that if the imputer model is uncongenial, in the sense that the model to create multiple imputations is a submodel to the one used by an analyst, then the multiple imputation estimators of the parameters in the analyst model could be biased (Meng 1994). In contrast, if the analyst model is a submodel of the one used by the imputer, then the multiple imputation estimators have no bias but are not efficient, even when the number of imputations is large. The confidence intervals are also conservative (that is, the actual coverage is larger than the nominal level).

The problem of uncongeniality of imputer/analyst models is probably more acute in this application than in the missing data problem. The entire synthetic sample is subject to model misspecification, whereas in the missing data context, only the imputed values, usually small relative to the number of observed values, are affected. Clearly, for the synthetic data approach to succeed, the model used to create synthetic populations (or equivalently, synthetic samples) should be expansive, in the sense that several models that may be fitted by analysts using the multiple synthetic samples should be submodels. Raghunathan et al. (2003) evaluated a nonparametric approach for creating multiple synthetic samples using an approximate Bayesian bootstrap (Rubin 1987). More work is needed in this direction. Creating synthetic data sets for longitudinal or panel surveys poses additional challenges. For example, suppose that D_1 is the data collected in wave 1 and the data collector has released the synthetic data sets $D_{\mathrm{syn},1}$ to the users. When the wave 2 data, D_2, are collected, then the creation of $D_{\mathrm{syn},2}$ poses a challenge because the data collector may not want to change $D_{\mathrm{syn},1}$. Although, statistically, one may want to generate from the predictive distribution, $\Pr(z_1, z_2|D_1, D_2)$, from the practical point of view, we may have to generate from $\Pr(z_2|D_{\mathrm{syn},1}, D_2)$. Further research is needed to investigate synthetic data generation methods for ongoing longitudinal studies.

The research community is used to accessing public-use data sets for their needs. The fact that the data sets released are synthetic may introduce suspicion. Therefore, it is important for data producers to develop trust with the ultimate users. It is incumbent upon both producer and user to come to terms with ultimate goal of protecting the respondents and performing valid statistical analyses.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Abowd JM, Stephens BW, Vilhuber L, Andersson F, McKinney KL, et al. 2009. The LEHD infrastructure files and the creation of the quarterly workforce indicators. In *Producer Dynamics: New Evidence from Micro Data*, ed. T Dunne, JB Jensen, MJ Roberts, pp. 149–230. Chicago: Univ. Chicago Press

Abowd JM, Woodcock SD. 2001. Disclosure limitation in longitudinal linked data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, ed. P Doyle, J Lane, J Theeuwes, L Zayatz, pp. 215–77. New York: North Holland

Abowd JM, Woodcock SD. 2004. Multiply-imputing confidential characteristics and file links in longitudinal linked data. In *Privacy in Statistical Databases*, ed. J Domingo-Ferrer, V Torra, pp. 290–97. Heidelberg, Ger.: Springer-Verlag

Bowen CM, Snoke J. 2019. Comparative study of differentially private synthetic data algorithms and evaluation standards. arXiv:1911.12704 [stat.AP]

Burman LE, Engler A, Khitatrakun S, Nunns JR, Armstrong S, et al. 2018. *Safely expanding research access to administrative tax data: creating a synthetic public use file and a validation server*. Res. Rep., Tax Policy Cent., Urban Inst., and Brookings Inst., Washington, DC

CMS (Cent. Medicare Medicaid Serv.). 2013. *Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)*. Baltimore, MD: US Cent. Medicare Medicaid Serv.

Cox LH. 1994. Matrix masking methods for disclosure limitation in micro data. *Surv. Methodol.* 20:165–69

Dalenius T. 1974. The invasion of privacy problem and statistics production—an overview. *Stat. Tidskr.* 3:213–25

Dalenius T. 1977. Privacy transformations for statistical information systems. *J. Stat. Plann. Inference* 1:73–86

Dalenius T. 1978. *Information privacy and statistics: a topical bibliography*. Rep. 41, Bur. Census, Dep. Commer., Washington, DC

DuMouchel W, Volinsky C, Johnson T, Cortes C, Pregibon D. 1999. Squashing flat files flatter. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. U Fayyad, S Chaudhuri, D Madigan, pp. 6–15. New York: ACM

Duncan GT, ed. 1993. Special issue: confidentiality and data access. *J. Off. Stat.* 9(2)

Duncan GT, Pearson RB. 1991. Enhancing access to micro-data while protecting confidentiality: prospects for the future. *Stat. Sci.* 6:219–39

Dwork C, McSherry F, Nissim K, Smith A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006*, ed. S Halevi, T Rabin, pp. 265–84. New York: Springer

Fienberg SE, Makov UE, Steele RJ. 1998. Disclosure limitation using perturbation and related methods for categorical data (with discussions). *J. Off. Stat.* 14:485–511

Fienberg SE, Willenborg LCRJ, eds. 1998. Special issue: disclosure limitation methods for protecting the confidentiality of statistical data. *J. Off. Stat.* 14(4)

Fuller WA. 1993. Masking procedures for microdata disclosure limitation. *J. Off. Stat.* 9:383–406

Gouweleeuw PK, Willenborg LCRJ, de Wolf PP. 1998. Post randomization for statistical disclosure control: theory and implementation. *J. Off. Stat.* 14:463–78

Jabine TB. 1993. Statistical disclosure limitation practices of United States statistical agencies. *J. Off. Stat.* 9:427–54

Keller WJ, Kooiman P, eds. 1992. Special issue: proceedings of the International Symposium on Statistical Disclosure Avoidance. *Stat. Neerl.* 46(1)

Kennickell AB. 1999. Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In *Statistical Data Protection*, ed. J Domingo-Ferrer, pp. 248–67. Luxembourg: Off. Off. Publ. Eur. Communities

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

Little RJA. 1993. Statistical analysis of masked data. *J. Off. Stat.* 9:407–26

Little RJA, Liu F, Raghunathan TE. 2004. Statistical disclosure techniques based on multiple imputation. In *Applied Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. A Gelman, XL Meng, pp. 141–52. New York: Wiley

Madigan D, Raghavan N, DuMouchel W, Nason M, Posse C, Ridgeway G. 2002. Likelihood-based data squashing: a modeling approach to instance construction. *Data Min. Knowl. Discov.* 6:173–90

McClure D, Reiter JP. 2012. Differential privacy and statistical disclosure risk measures: an investigation with binary synthetic data. *Trans. Data Privacy* 5(3):535–52

McSherry F, Talwar K. 2007. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, pp. 94–103. Piscataway, NJ: IEEE

Meng XL. 1994. Multiple imputation with uncongenial sources of input (with discussion). *Stat. Sci.* 9:538–74

Natl. Acad. Sci. Eng. Med. 2017. *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: Natl. Acad. Press

Nowok B, Raab GM, Dibben C. 2016. synthpop: Bespoke creation of synthetic data in R. *J. Stat. Softw.* 74(11):1–26

NRC (Natl. Res. Counc.). 1993. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2003. *Protecting Participants and Facilitating Social and Behavioral Sciences Research*. Washington, DC: Natl. Acad. Press

NRC (Natl. Res. Counc.). 2005. *Expanding Access to Research Data: Reconciling Risks and Opportunities*. Washington, DC: Natl. Acad. Press

Raghunathan TE, Berglund P, Solenberger PW. 2018a. *Multiple Imputation in Practice: With Examples Using IVEware*. Boca Raton, FL: CRC

Raghunathan TE, Berglund P, Solenberger PW, Van Hoewyk J. 2018b. *IVEware: imputation and variance estimation software user guide*, Version 0.3. Tech. Rep., Surv. Methodol. Progr., Surv. Res. Cent., Inst. Soc. Res., Univ. Mich., Ann Arbor. **http://www.isr.umich.edu/src/smp/ive/**

Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* 27:85–95

Raghunathan TE, Reiter JP, Rubin DB. 2003. Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* 19:1–16

Raghunathan TE, Rubin DB. 2000. *Bayesian multiple imputation to preserve confidentiality in public-use data sets*. Talk presented at ISBA 2000: The Sixth World Meeting of the International Society for Bayesian Analysis, Heronissos, Greece

Reiter JP. 2002. Satisfying disclosure restrictions with synthetic data sets. *J. Off. Stat.* 18:531–43

Reiter JP. 2003. Inferences for partially synthetic, public use microdata sets. *Surv. Methodol.* 29:181–88

Reiter JP, Dreschsler J. 2010. Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *Stat. Sin.* 20:405–21

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley

Rubin DB. 1993. Discussion of statistical disclosure limitation. *J. Off. Stat.* 9:461–68

Sakshaug JW, Raghunathan TE. 2010. Synthetic data for small area estimation. In *PSD 2010: Privacy in Statistical Databases*, ed. J Domingo-Ferrer, E Magkos, pp. 162–73. Heidelberg, Ger.: Springer

Surendra H, Mohan HS. 2017. A review of synthetic data generation methods for privacy preserving data publishing. *Int. J. Sci. Technol. Res.* 6:95–101

Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, et al. 2018. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Med. Inform. Assoc.* 25:230–38

Wasserman L, Zhou S. 2010. A statistical framework for differential privacy. *J. Am. Stat. Assoc.* 105(489):375–89

Annual Review of
Statistics and
Its Application

Volume 8, 2021

Contents

Modeling Player and Team Performance in Basketball
Zachary Terner and Alexander Franks 1

Graduate Education in Statistics and Data Science: The Why, When, Where, Who, and What
Marc Aerts, Geert Molenberghs, and Olivier Thas 25

Statistical Evaluation of Medical Tests
Vanda Inácio, María Xosé Rodríguez-Álvarez, and Pilar Gayoso-Diz 41

Simulation and Analysis Methods for Stochastic Compartmental Epidemic Models
Tapiwa Ganyani, Christel Faes, and Niel Hens 69

Missing Data Assumptions
Roderick J. Little 89

Consequences of Asking Sensitive Questions in Surveys
Ting Yan 109

Synthetic Data
Trivellore E. Raghunathan 129

Algorithmic Fairness: Choices, Assumptions, and Definitions
Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum 141

Online Learning Algorithms
Nicolò Cesa-Bianchi and Francesco Orabona 165

Space-Time Covariance Structures and Models
Wanfang Chen, Marc G. Genton, and Ying Sun 191

Extreme Value Analysis for Financial Risk Management
Natalia Nolde and Chen Zhou 217

Sparse Structures for Multivariate Extremes
Sebastian Engelke and Jevgenijs Ivanovs 241

Compositional Data Analysis
Michael Greenacre 271

Distance-Based Statistical Inference
Marianthi Markatou, Dimitrios Karlis, and Yuxin Ding 301

A Review of Empirical Likelihood
Nicole A. Lazar 329

Tensors in Statistics
Xuan Bi, Xiwei Tang, Yubai Yuan, Yanqing Zhang, and Annie Qu 345

Flexible Models for Complex Data with Applications
Christophe Ley, Slađana Babić, and Domien Craens 369

Adaptive Enrichment Designs in Clinical Trials
Peter F. Thall 393

Quantile Regression for Survival Data
Limin Peng 413

Statistical Applications in Educational Measurement
Hua-Hua Chang, Chun Wang, and Susu Zhang 439

Statistical Connectomics
Jaewon Chung, Eric Bridgeford, Jesús Arroyo, Benjamin D. Pedigo, Ali Saad-Eldin, Vivek Gopalakrishnan, Liang Xiang, Carey E. Priebe, and Joshua T. Vogelstein 463

Twenty-First-Century Statistical and Computational Challenges in Astrophysics
Eric D. Feigelson, Rafael S. de Souza, Emille E.O. Ishida, and Gutti Jogesh Babu 493

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics