Synthetic Data: A Look Back and A Look Forward

Jerome P. Reiter*

*Box 90251, Duke University, Durham, NC 27708, USA.

E-mail: jreiter@duke.edu

Abstract. When initially proposed, synthetic data for disclosure control was generally dismissed as unlikely to be implemented in practice. Thirty years later, synthetic data are becoming a staple of the disclosure limitation toolkit. We now see synthetic public use files for several major data products with more on the way. In this article, I review the progression of synthetic data, describe some unresolved challenges, and speculate on its future.

Keywords. disclosure; imputation; privacy; risk; verification.

1 Introduction

Almost 30 years ago, Don Rubin [1] published a radical proposal for statistical agencies seeking to share data with the public: disseminate public use files comprising simulated data. Today, Rubin's suggestion, now known as synthetic data, has taken root. Several statistical agencies and other data stewards, henceforth all called agencies, are turning to synthetic data strategies to create public use files from confidential data. As examples, the U.S. Census Bureau has released synthetic public use files for the Survey of Income and Program Participation [2] and Longitudinal Business Database [3], and it is currently overhauling the confidentiality protection methods for the American Community Survey (ACS), its flagship survey, to use synthetic data [4]. The U.S. Internal Revenue Service is making a synthetic public use file of individual tax returns [5]. The U.S. Agency for Healthcare Research and Quality has funded a project to create a synthetic version of a national healthcare claims database [6]. Outside the U.S., synthetic data have been used by Statistics Canada, Statistics New Zealand, the Australian Bureau of Statistics, the United Kingdom Office of National Statistics, the German Institute for Employment Research, and the Scottish Longitudinal Study, to name a few examples. Indeed, the United Nations Economic Commission for Europe recently authored a synthetic data manual for national statistics organizations [7].

One can understand the interest in synthetic data as a disclosure protection strategy; the approach has many appealing features. Most prominently, releasing synthetic data can reduce disclosure risks. Attackers may find it difficult to identify individuals who participated in the confidential data, and to learn the values of those individuals' sensitive attributes, when the released data are not actual, collected values. Furthermore, with appropriate data generation and estimation methods based on the concepts of multiple imputation [8, 9, 10, 11], the approach enables data users to make valid inferences for a variety of

estimands using standard, complete-data statistical methods and software. Other benefits are discussed, for example, in [12, 13, 14, 15, 16].

In this article, I review the progression of synthetic data since Rubin's original proposal and speculate on its future. I focus on synthetic data as a disclosure protection strategy for public use data files. Certainly, other uses of synthetic data have become prevalent, such as for training and testing machine learning algorithms and for facilitating business functions, but I do not cover these here. Throughout, I cite methodological and practical research on synthetic data; however, my selection of references is far from an exhaustive review of the literature on synthetic data. Many excellent works are not cited here. Finally, I am honored to write this article to mark the fifteenth anniversary of *Transactions on Data Privacy*. The journal has played a key role in advancing the practice of confidentiality protection including synthetic data, e.g., [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27].

2 Thirty Years of Synthetic Data

Readers of *Transactions on Data Privacy* know well the critical importance of protecting data subjects' confidentiality in public use files. Traditionally, agencies have attempted to do so by altering data before release [28, 29, 30]. These perturbation techniques are typically applied at low intensity to minimize degradation of the quality of secondary data analyses [31]. However, with the emergence of accessible digital data and powerful computational resources, attackers now can obtain huge amounts of information to use to link released records to external data files [32]. This concern has led several agencies to consider abandoning traditional statistical disclosure limitation methods in favor of newer methods like synthetic data.

Motivated by concerns over traditional methods, Rubin [1] proposed that agencies release multiply-imputed, synthetic data sets. In this approach, the agency (i) randomly and independently samples units from the sampling frame to comprise each synthetic data set, (ii) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these data sets to the public. A similar approach was suggested by [33]. These are now called fully synthetic data or also completely synthetic data [14].

In the same journal issue as Rubin's proposal, Little [34] proposed a variant of synthetic data, now called partially synthetic data [9] or incompletely synthetic data [14]. These comprise the units originally surveyed with only some collected values replaced with multiple imputations. For example, the agency might simulate sensitive variables or quasi-identifiers for units in the sample with rare combinations of quasi-identifiers [35, 36, 37, 38]; or, the agency might replace all data for selected sensitive variables or quasi-identifiers [12, 39, 40, 41]. The former strategy has been employed by the U.S. Federal Reserve Board in the Survey of Consumer Finances. They replace monetary values at high disclosure risk with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values [42]. It also is used by the U.S. Census Bureau to protect the identities of people in group quarters (e.g., prisons, shelters) in the American Community Survey [43]. They replace quasi-identifiers for records at high disclosure risk with imputations. The latter strategy has been employed by the U.S. Census Bureau to create the synthetic Survey of Income and Program Participation [2]. For this public use file, the Census Bureau synthesizes all values of around 600 variables, leaving just a handful at their original values.

The U.S. Census Bureau's Longitudinal Business Database (LBD), which contains the an-

nual total payroll and employee size since 1975 for every U.S. business establishment with paid employees, is an informative case study on how synthetic data can enable public access to otherwise confidential data [3, 44]. Data from the LBD derive from tax files compiled by the U.S. Internal Revenue Service (IRS), making the LBD subject to both Title 13 and Title 26 of the U.S. code. As a result, no actual values for individual establishments in the LBD can be released to the public; even the fact that an establishment filed taxes—and hence is in the dataset—is protected. Hence, top coding cannot be used on monetary data as a large fraction of exact values would be released. This also suggests that swapping would have to be done at an extremely high rate, in which case the released data would be useless for any analysis involving relationships with swapped variables. Furthermore, many variables of interest to researchers and policy makers, for example the number of employees and total payroll, have highly skewed distributions even within industry classifications. Hence, the amount of added noise necessary to disguise these observations would have to be very large, resulting in data of limited usefulness.

The only way that the Census Bureau and IRS were willing to release a public use version of the LBD was to make it synthetic [3]. The general strategy was to construct a joint distribution of all the variables using sequential conditional modeling, e.g., $f(x_1, x_2, x_3, \ldots, x_p) = f(x_1)f(x_2 \mid x_1)f(x_3 \mid x_1, x_2) \cdots f(x_p \mid x_1, \ldots, x_{p-1})$. The conditional models were tuned to different types of variables, such as logistic regressions for binary variables and linear regressions for continuous variables. The conditional specification also allowed the Census Bureau to conveniently incorporate logical relationships among the variables. The synthesis process was done separately for each industry type using models fit to the data for that industry, so as to allow for efficient and parallelizable computation. The synthetic LBD are available for download from a dedicated Census Bureau website [45].

The initial version of the synthetic LBD used parametric statistical models. However, researchers have developed a variety of synthetic data generators based on techniques from machine learning. For example, agencies can use classification and regression trees (CART) for the conditional modeling [46]. These are the primary engines for the Scottish Longitudinal Study data synthesis [47] and a leading contender for the American Community Survey data synthesis. Other examples of nonparametric conditional models include random forests [24] and support vector machines [48, 49]. Additionally, researchers have proposed deep learning methods like generative adversarial networks to generate synthetic data, e.g., [50, 51, 52]. These deep learning methods are multivariate synthesizers, in the sense that they generate all values at once rather than in a sequence of conditional models.

3 Unresolved Challenges: Disclosure Risk and Data Quality

Despite all the methodological advances and the existence of genuine applications of synthetic data, there are practical obstacles to broad-scale implementation of the approach. Here I review two of these obstacles: assessing disclosure risk and enabling analysts to assess the quality of synthetic data results.

3.1 Disclosure Risk Assessment for Fully Synthetic Data

Generally, it is thought that fully synthetic data should carry low risks of disclosures, since the released data do not correspond to actual records. This largely eliminates the kinds of

record linkage attacks that have broken typical disclosure control methods, as it is nonsensical for attackers to match synthetic records to external files [53]. However, this does not mean there are zero risks. To illustrate with a simple example, suppose that the confidential data comprise four binary variables, and only one individual in the confidential data has a value of $(0, 0, 0, 0)$. Suppose an agency considers generating synthetic data by sampling from a multinomial distribution with probabilities equal to the empirical frequencies in the 16 cells. When the synthetic data include a case at $(0, 0, 0, 0)$, the attacker knows that someone in the confidential data must have those values, which could be a disclosure risk. Related, an agency's synthesis model may perfectly predict some variable for a certain type of individual, so that synthetic values of this variable always match the actual values for this type of individual. The agency may consider these unacceptable disclosure risks.

While there is need to examine disclosure risks in synthetic data, there is no standard for doing so, especially in fully synthetic data. Instead, disclosure checks tend to be *ad hoc*. For example, in the synthetic LBD, the Census Bureau examined whether large synthetic employment values in specific years are too close to confidential values. They also examined distributions of actual start years for each synthetic start year value to ensure that there is sufficient diversity in the mapping from one to the other. All of these checks proceed one variable at a time, assume attackers with no external knowledge, and do not provide quantitative measures of risk. Related notions of disclosure risk are suggested in [54].

There are germs of principled approaches in the literature. For partially synthetic data, [55] and [56] present risk measures for an attacker who knows the collected values of a single target record and searches the released data to identify that record. These approaches do not apply to fully synthetic data and, even for partially synthetic data, assume the attacker has access to a small number of variables on the target. One can compute probabilities that attackers correctly guess individuals' sensitive values given the synthetic data [57]; related examples and approaches are in [16, 40, 41, 58, 59, 60]. However, all of these examples use data with a small number of variables. Computing these probabilities for high-dimensional synthetic data is a computationally intensive and challenging task, and as such has never been applied in production settings. Moving these ideas from theory to practice is an important research objective.

An alternative approach that is feasible in some settings is to generate synthetic data to satisfy, at least approximately, some variant of differential privacy, e.g., as in [27, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71]. This is done, for example, by the U.S. Census Bureau to construct data products from the 2020 population census. To date, however, there are no differentially private algorithms that can generate synthetic data with low errors and high privacy protection for high-dimensional surveys with complex sampling designs and the typical edit-imputation and re-weighting procedures for nonresponse and calibration. This is the primary reason why the Census Bureau, which has a publicly stated goal of moving towards formal privacy for its public use data products, is using model-based synthetic data for the re-design of the confidentiality protection methods for the ACS.

3.2 Enabling Quality Checks with Verification Measures

The quality of synthetic data are only as good as the quality of the synthesis models [16]. Any relationships or distributional features not present in the synthesis models also are not present in the synthetic data. Thus, agencies should thoroughly check synthetic data before releasing them to the public. For example, in the synthetic LBD, the Census Bureau examined various statistics in the confidential and synthetic data, such as longitudinal trends in total annual employment, the number of jobs created each year, and the total payroll by

industry. These analyses were selected by subject matter experts at the Census Bureau.

Specific comparisons are useful and necessary, but they offer only so much comfort to users interested in other analyses. How does such a user know whether their results from the synthetic data are any good? Inevitably, some synthetic data inferences will deteriorate significantly because of imperfect synthesis models. Thus, it is arguably essential that agencies develop ways to provide feedback to users about the quality of inferences from synthetic data for specific estimands.

One proposed solution is to build a verification server [53, 72, 73]. The basic idea is as follows. The analyst, who has access only to the synthetic data, submits a query to the verification server for the results of a statistical model; for example, the coefficients in a regression or the mean of some variable in a subpopulation. The server, which has both the confidential and synthetic data, performs the analysis on both data sources. From the results, the server calculates analysis-specific measures of the fidelity of one to the other. For example, when the query is a regression coefficient, one verification measure is the overlap of the 95% confidence intervals for the coefficient when computed with the confidential data and with the synthetic data [74]. The server returns the value of the verification measure to the analyst. If the analyst feels that the intervals overlap adequately, the synthetic data have high utility for their analysis. With such feedback, analysts can be confident about results with good quality and avoid publishing results with poor quality [75].

There is a catch, however: verification measures also leak information about the confidential data that could be used for disclosure attacks [17, 72]. Clever attackers could submit queries for verification that, perhaps in combination with other information, allow them to estimate confidential values too accurately. Thus, a key area of research is to develop methods for verification that allow agencies to control this additional information leakage, e.g., by designing differentially private verification measures like those in [53].

A closely related option is for the agency to run the user's statistical analysis on the confidential data and report back a disclosure-protected version of the results. This is known as validation (as opposed to verification) of results. In fact, the Census Bureau allows users to submit their programs for validation with the synthetic Survey of Income and Program Participation and the synthetic LBD [76]. Validation also leaks information about the confidential data, further highlighting the importance of continued research on ways to provide users feedback on the quality of their synthetic data inferences.

4 The Future of Synthetic Data (?)

What does the future have in store for synthetic data? In this section, I offer a few thoughts on this question.

Any discussion of the future of data dissemination strategies should begin with speculation about how agencies might define data privacy and confidentiality (P&C) going forward. Traditionally, agencies have focused assessments of disclosure risks on whether individuals can be re-identified in a database. Indeed, many laws and regulations around data sharing explicitly speak to re-identifications. For many government surveys and censuses, however, it is not clear (to me) why the fact that someone is in the database needs to be confidential. For example, does someone only knowing that I participated in the ACS have any implication for my well being? I think not. Of course, learning that an individual is a member of a database can be problematic in some contexts, for example, a survey of people with a specific health condition. However, protection from re-identification risk may not need be the default priority that it is today. Instead, agencies might focus on reduc-

ing the risks of attribute disclosures and be less concerned with preventing re-identification disclosures.

For statistical agencies, one particularly relevant and recent shift in disclosure limitation practice is the desire to formalize P&C mathematically. This is evident in the invention of differential privacy and its variants. Differential privacy defines disclosure risks as relative; that is, how much information attackers can learn about any individual from the released statistics when that individual's data are used to compute the statistics compared to when those data are not used. This is a significant change from existing disclosure risk paradigms, which tend to focus on absolute risks to individuals in the data. This perspective that disclosure risks are relative naturally leads to the question, what information should be considered known at baseline? Attackers already can learn massive amounts of information on many of us with minimal effort. As examples, they can search social media and commercial websites to glean demographic and housing information, and they can purchase information about individuals from companies whose business model is to sell data. In the near future, it is conceivable that all individuals' factual information (comprising, e.g., certain demographic variables) might be considered already known to attackers when evaluating relative disclosure risks.

How might such a consideration impact applications of synthetic data? For agencies that let go of protecting against re-identifications and focus on protecting against disclosures of sensitive attributes, a natural approach is to release partially synthetic data. These could comprise the sampled individuals with their factual variables (that are considered publicly available) kept at their collected values and their sensitive attributes replaced with simulated values. This should make it easier for agencies to generate synthetic data with high analytic validity, as it reduces reliance on synthesis models. Further, an emphasis on attribute disclosure may facilitate generation of synthetic data that satisfy new mathematical definitions of privacy, for example, by using frameworks like Pufferfish [77]. Finally, by letting go of re-identification disclosures, it may be easier for agencies to design useful verification (or validation) measures that introduce low additional risks of disclosure.

Societal expectations of privacy are changing and are likely to continue to change in the future. Data dissemination policies of course must adapt to such changes. Thirty years of synthetic data have seen thirty years of innovation. I am excited to see what the next thirty years has in store for us.

References

[1] D. B. Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:462–468, 1993.

[2] J. Abowd, M. Stinson, and G. Benedetto. Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program, 2006.

[3] S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd. Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review*, 79:363–384, 2011.

[4] M. H. Freiman, A. D. Lauger, and J. P. Reiter. Formal privacy and synthetic data for the American Community Survey. Technical report, United States Bureau of the Census, 2018.

[5] L. E. Burman, A. Engler, S. Khitatrakun, J. R. Nunns, S. Armstrong, J. Iselin, G. MacDonald, and P. Stallworth. Safely expanding research access to administrative tax data: Creating a

synthetic public use file and a validation server. https://www.irs.gov/pub/irs-soi/18resconburman.pdf. Accessed: 2022-01-12.

[6] Agency for Healthcare Research and Quality. Synthetic Healthcare Database for Research (SyH-DR). https://www.ahrq.gov/data/syh-dr.html. Accessed: 2022-01-12.

[7] United Nations Economic Commission for Europe. Synthetic Data for National Statistical Organizations. https://statswiki.unece.org/display/SDS/Synthetic+Data+Sets+public?preview=%2F282330193%2F330369384%2FHLG-MOS+Synthetic+Data+Guide.docx. Accessed: 2022-01-12.

[8] T. E. Raghunathan, J. P. Reiter, and D. B. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1–16, 2003.

[9] J. P. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–189, 2003.

[10] J. P. Reiter. Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131:365–377, 2005.

[11] J. P. Reiter and T. E. Raghunathan. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102:1462–1471, 2007.

[12] J. M. Abowd and S. D. Woodcock. Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, pages 290–297. New York: Springer-Verlag, 2004.

[13] J. Drechsler. *Synthetic Datasets for Statistical Disclosure Control*. New York: Springer, 2011.

[14] G. M. Raab, B. Nowok, and C. Dibben. Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, 7:67–97, 2017.

[15] J. P. Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18:531–544, 2002.

[16] J. P. Reiter. Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168:185–205, 2005.

[17] D. McClure and J. P. Reiter. Differential privacy and statistical disclosure risk measures: An illustration with binary synthetic data. *Transactions on Data Privacy*, 5:535–552, 2012.

[18] G. Amitai and J. P. Reiter. Differentially private posterior summaries for linear regression coefficients. *Journal of Privacy and Confidentiality*, 8:Article 2, 2018.

[19] J. Drechsler, S. Bender, and S. Rässler. Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy*, 1:105–130, 2008.

[20] R. Hornby and J. Hu. Identification risks evaluation of partially synthetic data with the identification risk calculation R package. *Transactions on Data Privacy*, 14:37–52, 2021.

[21] J. Hu. Bayesian estimation of attribute and identification disclosure risks in synthetic data. *Transactions on Data Privacy*, 12:61–89, 2019.

[22] A. Oganian and J. Domingo-Ferrer. Local synthesis for disclosure limitation that satisfies probabilistic k-anonymity criterion. *Transactions on Data Privacy*, 10:61–81, 2017.

[23] Y. Park and J. Ghosh. PeGS: Perturbed Gibbs samplers that generate privacy-compliant synthetic data. *Transactions on Data Privacy*, 7:253–282, 2014.

[24] G. Caiola and J. P. Reiter. Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, 3:27–42, 2010.

[25] Jennifer Taub, Mark Elliot, and Joseph W. Sakshaug. The impact of synthetic data generation on data utility with application to the 1991 UK samples of anonymised records. 13:1–23, 2020.

[26] H. Yu and J. P. Reiter. Differentially private verification of regression predictions from synthetic data. *Transactions on Data Privacy*, 11:279–297, 2018.

[27] F. Liu. Model-based differentially private data synthesis and statistical inference in multiple synthetic datasets. *Transactions on Data Privacy*, 15:141–175, 2022.

[28] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholdt, K. Spicer, and P-P de Wolf. *Statistical Disclosure Control*. Chichester: Wiley, 2012.

[29] A. F. Karr. Data sharing and data access. *Annual Review of Statistics and Its Application*, 3:113–132, 2016.

[30] L. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*. New York: Springer-Verlag, 2001.

[31] J. P. Reiter. Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly*, 76:163–181, 2012.

[32] J. P. Reiter. Differential privacy and federal data releases. *Annual Review of Statistics and Its Applications*, 6:85–101, 2019.

[33] S. E. Fienberg. A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical report, Department of Statistics, Carnegie-Mellon University, 1994.

[34] R. J. A. Little. Statistical analysis of masked data. *Journal of Official Statistics*, 9:407–426, 1993.

[35] R. J. A. Little, F. Liu, and T. E. Raghunathan. Statistical disclosure techniques based on multiple imputation. In A. Gelman and X. L. Meng, editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 141–152. New York: John Wiley & Sons, 2004.

[36] D. An and R. J. A. Little. Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A*, 170:923–940, 2007.

[37] J. Drechsler and J. P. Reiter. Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105:1347–1357, 2010.

[38] J. Drechsler and J. P. Reiter. Combining synthetic data with subsampling to create public use microdata files for large scale surveys. *Survey Methodology*, 38:73–79, 2012.

[39] J. M. Abowd and S. D. Woodcock. Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, editors, *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 215–277. Amsterdam: North-Holland, 2001.

[40] T. Paiva, A. Chakraborty, J. P. Reiter, and A. E. Gelfand. Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in Medicine*, 33:1928–1945, 2014.

[41] H. Wang and J. P. Reiter. Multiple imputation for sharing precise geographies in public use data. *Annals of Applied Statistics*, 6:229–252, 2012.

[42] A. B. Kennickell. Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, editors, *Record Linkage Techniques, 1997*, pages 248–267. Washington, D.C.: National Academy Press, 1997.

[43] S. Hawala. Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association, 2008.

[44] S. K. Kinney and J. P. Reiter. Making public use, synthetic files of the Longitudinal Business Database. In *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association, 2007.

[45] United States Bureau of the Census. Synthetic Longitudinal Business Database (SynLBD). `https://www.census.gov/programs-surveys/ces/data/public-use-data/synthetic-longitudinal-business-database.html`. Accessed: 2022-01-12.

[46] J. P. Reiter. Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21:441–462, 2005.

[47] B. Nowok, G. M. Raab, and C. Dibben. synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, 74:1–26, 2016.

[48] J. Drechsler. Using support vector machines for generating synthetic datasets. In J. Domingo-Ferrer and E. Magkos, editors, *Privacy in Statistical Databases*, pages 148–161. New York: Springer-Verlag, 2010.

[49] J. Drechsler and J. P. Reiter. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*, 55:3232–3243, 2011.

[50] E. Choi, S. Biswal, B. Malin, J. Duke, W. Stewart, and J. Sun. Generating multi-label discrete patient records using generative adversarial networks. *Proceedings of Machine Learning Research*, 68:286–305, 2017.

[51] I. Kaloskampis, C. Joshi, C. Cheung, D. Pugh, and L. Nolan. Synthetic data in the civil service. *Significance*, 17:18–23, 2020.

[52] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11:1071–1083, 2018.

[53] A. F. Barrientos, A. Bolton, T. Balmat, J. P. Reiter, J. M. de Figueiredo, A. Machanavajjhala, Y. Chen, C. Kneifel, and M. DeLong. Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *Annals of Applied Statistics*, 12:1124–1156, 2018.

[54] J. Taub, M. Elliot, M. Pampaka, and D. Smith. Differential correct attribution probability for synthetic data: An exploration. In J. Domingo-Ferrer and F. Montes, editors, *Privacy in Statistical Databases*, pages 122–137. 2018.

[55] J. Drechsler and J. P. Reiter. Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin, editors, *Privacy in Statistical Databases*, pages 227–238. New York: Springer-Verlag, 2008.

[56] J. P. Reiter and R. Mitra. Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, 1:99–110, 2009.

[57] J. P. Reiter, Q. Wang, and B. Zhang. Bayesian estimation of disclosure risks in multiply imputed, synthetic data. *Journal of Privacy and Confidentiality*, 6:Article 2, 2014.

[58] J. Hu, J. P. Reiter, and Q. Wang. Disclosure risk evaluation for fully synthetic data. In J. Domingo-Ferrer, editor, *Privacy in Statistical Databases*, pages 185–199. Heidelberg: Springer, 2015.

[59] D. Manrique-Vallier and J. Hu. Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society, Series A*, 181:635–647, 2018.

[60] D. McClure and J. P. Reiter. Assessing disclosure risks for synthetic data with arbitrary intruder knowledge. *Statistical Journal of the International Association of Official Statistics*, 32:109–126, 2016.

[61] J. Abowd and L. Vilhuber. How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygun, editors, *Privacy in Statistical Databases*, pages 239–246. New York: Springer-Verlag, 2008.

[62] C. M. Bowen and F. Liu. Comparative study of differentially private data synthesis methods. *Statistical Science*, 35:280–307, 2020.

[63] E. Bao, X. Xiao, J. Zhao, D. Zhang, and B. Ding. Synthetic data generation with differential privacy via Bayesian networks. *Journal of Privacy and Confidentiality*, 11:Article 4, 2021.

[64] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the 27th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems*, pages 273–282, 2007.

[65] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th ACM SIGACT Symposium on Theory of Computing*, pages 609–618, 2008.

[66] C. M. Bowen and J. Snoke. Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge. *Journal of Privacy and*

Confidentiality, 11:Article 4, 2021.

[67] A. S. Charest. How can we analyze differentially private synthetic datasets. *Journal of Privacy and Confidentiality*, 2:2:Article 3, 2010.

[68] V. Karwa and A. S. Slavkovic. Differentially private graph degree sequences and synthetic graphs. In J. Domingo-Ferrer and I. Tinnirello, editors, *Privacy in Statistical Databases, Lecture Notes in Computer Science 7556*, pages 273–285. Berlin: Springer-Verlag, 2012.

[69] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering*, pages 277–286, 2008.

[70] R. McKenna, G. Miklau, and D. Sheldon. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *Journal of Privacy and Confidentiality*, 11:Article 4, 2021.

[71] Y. Rinott, C. M. O'Keefe, N. Shlomo, and C. Skinner. Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Science*, 33:358–385, 2018.

[72] J. P. Reiter, A. Oganian, and A. F. Karr. Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis*, 53:1475–1482, 2009.

[73] A. F. Karr and J. P. Reiter. Using statistics to protect privacy. In J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, editors, *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, pages 276–295. 2014.

[74] A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60:224–232, 2006.

[75] J. P. Reiter and J. Drechsler. Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, 20:405–422, 2010.

[76] L. Vilhuber, J. M. Abowd, and J. P. Reiter. Synthetic establishment microdata around the world. *Statistical Journal of the International Association for Official Statistics*, 32:65–68, 2016.

[77] D. Kifer and A. Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39:1–36, 2014.