American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Synthetic Data Governance for Collaborative Social Research

Dr. Amelia J. Foster

Associate Professor, University of Manchester, United Kingdom

Abstract

Collaborative social research increasingly depends on access to administrative, survey, population, behavioral, and linked datasets that may contain confidential or personally sensitive information. Conventional restrictions on microdata access can protect participants but may simultaneously constrain multidisciplinary collaboration, methodological experimentation, reproducibility, researcher training, and cross-institutional innovation. Synthetic data offer a potentially valuable intermediary by generating artificial records from statistical or computational models designed to preserve selected properties of protected source data without directly reproducing the complete original records. The Office for National Statistics recognizes synthetic data as potentially useful for research, testing, processing, and statistical purposes when real data are inaccessible, while explicitly warning that synthetic datasets do not necessarily reproduce all properties of source data and require disclosure assessment before wider release. The UK Information Commissioner's Office similarly emphasizes that synthetic data may or may not be anonymous and that privacy risk should be assessed rather than assumed away merely because records are synthetic.

This study develops a governance framework for synthetic data used in collaborative social research. The paper integrates privacy and disclosure control, statistical utility, representativeness, provenance, purpose limitation, researcher accountability, reproducibility, and cross-institutional access governance. A conceptual-methodological design is combined with a transparent simulation-based assessment of six governance-risk domains. The analysis identifies privacy and disclosure risk as the largest simulated contribution to overall governance exposure, followed by representativeness and bias, provenance and documentation, secondary use, purpose alignment, and reproducibility. These values are explicitly illustrative rather than empirical estimates. The study proposes a governance architecture that combines synthetic-data classification, use-case specification, privacy testing, utility validation, dataset documentation, controlled access, research-ethics oversight, human accountability, and validation against protected source data before consequential findings are published. The paper concludes that synthetic data should not be treated as inherently anonymous, inherently representative, or automatically suitable for unrestricted open science. Their value lies instead in enabling proportionate and auditable research access when technical protections are integrated with institutional governance.

Keywords: synthetic data, social research, data governance, privacy-preserving data sharing, research collaboration, statistical disclosure control, data utility, reproducibility, research ethics, differential privacy

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

1. Introduction

Contemporary social research increasingly depends on data that are both scientifically valuable and institutionally difficult to share. Administrative records, longitudinal surveys, welfare records, educational data, employment histories, justice-system information, health-linked social datasets, and digitally generated behavioral records can support powerful analyses of inequality, mobility, public policy, social vulnerability, and institutional outcomes. At the same time, these resources may contain direct identifiers, indirect identifiers, sensitive attributes, rare combinations of characteristics, or information capable of becoming disclosive when linked with external sources. Collaborative projects involving universities, government departments, research infrastructures, and international partners therefore face a persistent tension between expanding legitimate scientific access and maintaining confidentiality. The Five Safes framework has been widely used to structure such decisions around safe projects, people, settings, data, and outputs, although recent scholarship also emphasizes that the framework should continue evolving rather than becoming a static compliance checklist.

Synthetic data have emerged as one response to this tension. Rather than distributing the protected source records themselves, a data custodian estimates a statistical or computational model and generates artificial observations intended to reproduce selected structures of the underlying dataset. Synthetic data can be fully synthetic, where all relevant values are generated, or partially synthetic, where only selected attributes or records are replaced. NIST defines synthetic data generation as a process in which seed data are used to create artificial data possessing some of the statistical characteristics of those seed data. The ONS similarly treats synthetic data as potentially useful substitutes for research and development where real data cannot readily be accessed, but stresses that their fitness depends on the purpose for which they were produced.

The apparent attractiveness of synthetic data can nevertheless encourage an overly simple assumption: because the records are artificial, the privacy problem has disappeared. That inference is unsafe. The Information Commissioner's Office expressly notes that synthetic data generated from models of original data may or may not be anonymous. Its privacy-enhancing-technology guidance further explains that higher resemblance to real data may increase utility while potentially increasing the likelihood that sensitive information is revealed. Stadler, Oprisanu, and Troncoso reached a similarly cautionary conclusion in their empirical evaluation, showing that synthetic data do not automatically provide a superior privacy–utility trade-off and may remain susceptible to inference risks.

A second challenge concerns scientific validity. Collaborative researchers may use synthetic datasets for exploratory analysis, code development, teaching, protocol testing, statistical modelling, or—in some circumstances—substantive analysis. Yet a synthetic dataset preserves only those relationships represented sufficiently well by its generative mechanism. Snoke and colleagues distinguish general utility, referring to broader similarity between synthetic and original data distributions, from specific utility, referring to similarity of results for a particular analysis. This distinction is important because a dataset that is useful for one regression model may be unsuitable for another research question involving interactions, rare groups, nonlinear relationships, or subgroup inequalities. Recent work on terminology likewise shows that synthetic-data concepts such as utility, fidelity, privacy, and related terms remain inconsistently used, increasing the possibility of misunderstanding among researchers, policymakers, data providers, and the public.

2. Objectives of the Study

2.1 To Examine the Governance Risks of Synthetic Data Sharing

The first objective is to identify the principal governance risks that arise when synthetic datasets derived from protected social data are shared across researchers, institutions, or jurisdictions. The study examines privacy and disclosure, representativeness, model-induced bias, provenance, secondary use, purpose limitation, methodological validity, and responsibility for downstream research claims. Particular attention is given to the misconception that synthetic data are automatically anonymous. Current ICO guidance explicitly rejects such an assumption and requires attention to whether individuals remain identifiable and what could be learned if identification were successful.

2.2 To Develop a Collaborative Synthetic-Data Governance Framework

The second objective is to construct a governance model that combines technical privacy safeguards with research-governance mechanisms. The proposed model integrates dataset classification, project approval, researcher authorization, secure environments, privacy evaluation, utility assessment, provenance documentation, disclosure review, and validation against protected source data. This approach is informed by the Five Safes tradition but treats synthetic data as one control within a wider risk-management system rather than as a substitute for all other safeguards.

2.3 To Evaluate the Balance Between Access, Privacy, and Research Utility

The third objective is to demonstrate how competing governance risks can be assessed systematically. A simulation-based analysis is therefore used to distribute an illustrative composite governance risk across six domains. The numerical values are methodological examples rather than observations from real data-sharing projects. The analysis is intended to demonstrate how institutions could document governance priorities before releasing synthetic datasets for collaborative research.

3. Review of Literature

3.1 Synthetic Data, Privacy, and Disclosure Risk

Synthetic microdata have a substantial methodological history in statistical disclosure limitation. Reiter's early empirical work demonstrated how multiply imputed fully synthetic public-use microdata could be generated and evaluated for both inferential validity and confidentiality protection. The underlying principle is appealing: rather than modifying every original record through conventional masking, the data custodian generates new records from an estimated model. The synthetic observations can then support analysis without directly providing researchers with the protected microdata from which the synthesis model was learned.

This architecture, however, does not create a universal privacy guarantee. Stadler, Oprisanu, and Troncoso systematically evaluated modern generative approaches and showed that synthetic data may either remain vulnerable to inference attacks or suffer substantial utility degradation. Their findings challenge the treatment of synthetic data as a privacy "silver bullet" and demonstrate that disclosure protection must be tested empirically or supported by formal privacy guarantees rather than inferred from the label "synthetic."

Differential privacy provides a more formal approach to controlling information leakage. NIST's 2021 guidelines describe differential privacy as a mathematical framework for quantifying privacy loss and stress that real-world implementations require careful attention to algorithmic correctness, threat models, access controls, query models, and potential implementation hazards. Formal privacy parameters should therefore be considered part of an operational system rather than treated as a complete governance solution in isolation.

For collaborative social research, the implication is clear: the level of access permitted to a synthetic dataset should depend upon demonstrated disclosure properties and the context of use. An internally generated dataset accessed by accredited researchers within a controlled environment does not require the same residual-risk profile as a dataset placed on an unrestricted public repository. Governance should therefore evaluate privacy and access together rather than treating "synthetic" as a binary category of safety.

3.2 Statistical Utility, Fidelity, and Representativeness

The value of synthetic data depends on whether they preserve the properties required for the intended research. Snoke, Raab, Nowok, Dibben, and Slavković make a particularly useful distinction between general and specific utility. General utility concerns the broad distributional similarity of synthetic and original data, whereas specific utility considers whether a particular analysis produces sufficiently similar results in the two datasets. Their work develops propensity-score-based utility measures and demonstrates that a synthetic dataset may need to be evaluated using multiple complementary criteria rather than a single global score.

This distinction is fundamental for social research because research questions are frequently sensitive to minority groups, intersectional identities, rare events, geographic detail, nonlinear relationships, and complex longitudinal dependencies. A synthesis procedure that preserves marginal averages can still distort precisely the social inequalities that researchers intend to study. The ICO consequently advises organizations to detect and correct bias in synthetic-data generation and to consider whether the resulting data are representative, particularly when synthetic data may inform decisions with consequences for individuals.

Frayling and colleagues' 2021 review further highlights the lack of consistent terminology surrounding synthetic data, particularly concepts such as fidelity and utility. Their analysis recommends explicitly defining the type of synthetic data involved, connecting utility measures to intended use cases, and developing multidimensional evaluation frameworks. This is especially relevant to cross-institutional research because collaborators can otherwise assume that "high-fidelity synthetic data" support analytical uses that were never validated by the data producer.

3.3 Collaborative Governance, Transparency, and the Five Safes

Collaborative research introduces risks that cannot be managed by statistical synthesis alone. Researchers differ in training, project objectives, institutional controls, technical environments, publication plans, and access to external datasets. A synthetic dataset that poses low disclosure risk in one environment could become more revealing when linked with additional information or used by collaborators operating under weaker controls.

The Five Safes framework provides a useful starting point because it treats safe access as a combined function of projects, people, settings, data, and outputs. The UK Data Service applies the framework to research access to sensitive data, while Green and Ritchie describe its continuing evolution and the need to adapt the model to changing research environments.

The UK Statistics Authority's ethical guidance provides a complementary transparency-oriented approach. It recommends communicating the intended purpose of a synthetic dataset, how it was generated, whether it is fully or partially synthetic, its level of fidelity, and its limitations. Researchers publishing results should explain why synthetic data were used, why the dataset was appropriate for the research question, which limitations were encountered, and what tools or methods were applied.

4. Research Methodology

4.1 Research Design

The study adopts a conceptual-methodological design with simulation-based governance analysis. No identifiable records, survey respondents, confidential administrative datasets, or real collaborative research projects were analyzed. The numerical values reported in the Analysis section are synthetic values generated specifically to demonstrate the proposed framework.

The conceptual evidence base combines statistical literature on synthetic microdata, contemporary privacy research, official guidance on synthetic-data governance, differential-privacy guidance, ethical guidance, and research-access frameworks. The methodological emphasis is not on evaluating a particular data-generation algorithm but on defining the institutional conditions required before synthetic social data are produced, shared, analyzed, or used to support published claims. The literature reviewed above demonstrates that privacy protection, statistical fidelity, and research usability are related but distinct properties.

4.2 Governance Assessment Dimensions

Six domains were selected for the synthetic-data governance assessment: privacy and disclosure, representativeness and bias, provenance and documentation, secondary use and function creep, consent and purpose alignment, and reproducibility and validity. These domains synthesize the principal concerns emphasized across official guidance and methodological research. The model deliberately avoids treating privacy as the only objective because a dataset can protect confidentiality while still generating misleading or inequitable research.

Table 1: Proposed Governance Framework for Collaborative Synthetic Social Data

Governance Domain	Core Governance Question	Required Evidence	Recommended Control
Privacy and Disclosure	Could source individuals or sensitive attributes be inferred from the synthetic release?	Disclosure testing, attack-based evaluation, formal privacy information where applicable	Tiered access, privacy audit, disclosure review

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Governance Domain	Core Governance Question	Required Evidence	Recommended Control
Representativeness and Bias	Does synthesis preserve or distort socially important populations and relationships?	Subgroup utility, distributional comparisons, rare-category evaluation	Bias audit and subgroup validation
Provenance and Documentation	Can collaborators reconstruct how the synthetic dataset was created?	Source description, synthesis method, versioning, software and parameter documentation	Synthetic-data datasheet and version control
Secondary Use and Function Creep	Could the dataset be reused for purposes not considered during generation?	Defined permitted-use statement and downstream-use analysis	Data-use agreement and purpose restrictions
Consent and Purpose Alignment	Is synthesis consistent with the ethical and institutional basis under which source data were collected or accessed?	Ethics review, legal basis, public-interest assessment where applicable	Governance approval before generation and sharing
Reproducibility and Validity	Can findings obtained from synthetic data be reproduced and, where necessary, validated against protected source data?	Code, utility metrics, gold-standard validation protocol	Reproducible workflow and controlled final validation

Source: Author-developed framework informed by ONS synthetic-data policy, ICO PET guidance, UK Statistics Authority ethical guidance, Five Safes scholarship, and synthetic-data utility literature.

4.3 Simulation and Analytical Procedure

The simulation assigns a total governance-risk contribution of 100 points across the six domains. The purpose is not to estimate real-world frequencies but to demonstrate how a research consortium might prioritize controls during a governance review. Privacy and disclosure were assigned the largest illustrative contribution because residual disclosure can directly undermine the rationale for synthetic release. Representativeness was assigned the second-largest contribution because distorted population structures can compromise substantive social inference even when confidentiality is protected.

The analytical workflow proceeds from use-case definition to source-data governance, synthesis, privacy evaluation, utility testing, collaborative-access classification, research use, publication review, and—where substantive claims depend heavily on the synthetic data—validation against the original protected data. This "synthetic first, protected validation later" model is compatible with earlier

synthetic-data work in which researchers use synthetic datasets for exploratory analysis and subsequently conduct controlled gold-standard analysis against the original source.

5. Analysis

5.1 Simulated Governance-Risk Distribution

The first analytical stage evaluates the relative contribution of six governance domains.

Table 2: Simulated Governance-Risk Profile for Collaborative Synthetic-Data Research

Governance-Risk Domain	Simulated Contribution	Relative Priority	Primary Governance Failure
Privacy / Disclosure	26%	Very High	Synthetic records preserve information enabling inference or disclosure
Representativeness / Bias	21%	Very High	Minority groups or important social relationships are distorted
Provenance / Documentation	17%	High	Collaborators cannot reconstruct generation decisions or limitations
Secondary Use / Function Creep	14%	High	Data are reused beyond the purposes for which synthesis was evaluated
Consent / Purpose Alignment	12%	Moderate–High	Synthetic use diverges from ethical or institutional expectations
Reproducibility / Validity	10%	Moderate–High	Published results cannot be verified against the protected source
Total	100%	—	Composite illustrative governance exposure

Note: These percentages are simulated and are not prevalence estimates from real projects.

Privacy and disclosure account for the largest simulated contribution at 26%. This prioritization reflects the fact that synthetic generation does not itself establish anonymity. The ONS requires disclosure checking where synthetic data are to be shared more widely than the source data, while the ICO explicitly states that synthetic information may remain personal data depending on identifiability.

Representativeness and bias contribute 21%. This category is particularly important for social research because systematic distortion of small or disadvantaged groups can produce methodologically polished but substantively misleading conclusions. Synthetic-data governance should therefore examine subgroup-level utility rather than relying exclusively on overall similarity measures. The ICO recommends detecting and correcting bias and assessing whether generated data remain representative for their intended use.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

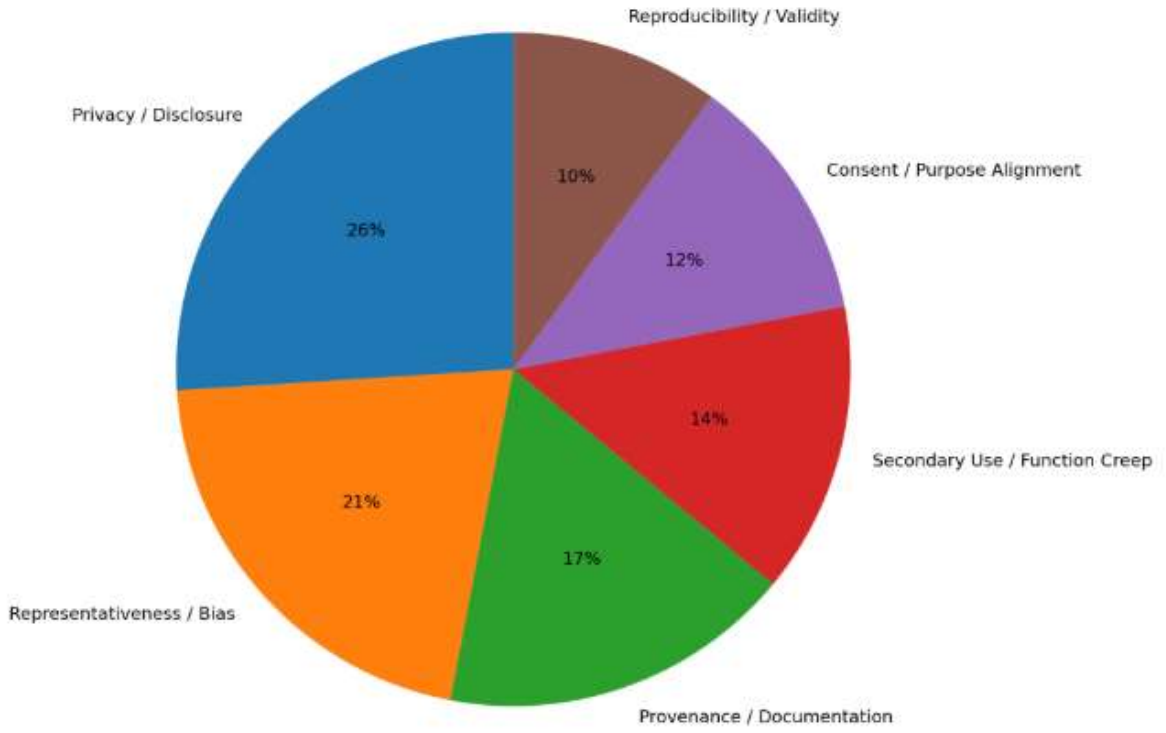
E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Provenance and documentation account for 17%. In collaborative environments, weak documentation can create divergence between what data producers know about the synthesis procedure and what downstream analysts assume. Frayling and colleagues identify inconsistent terminology as a continuing governance challenge, while the UK Statistics Authority recommends documenting dataset type, methods, fidelity, intended use, and limitations.

Secondary use contributes 14%, consent and purpose alignment 12%, and reproducibility and validity 10%. These lower simulated proportions should not be interpreted as indicating that these domains are unimportant. Rather, they illustrate the distinction between probability or frequency of risk and normative severity. A relatively uncommon validity failure can have substantial consequences if synthetic evidence becomes the basis of a major public-policy claim.

5.2 Graphical Analysis

Graph 1: Simulated Distribution of Governance Risks in Collaborative Synthetic-Data Research



The pie graph demonstrates that governance responsibility is distributed rather than concentrated in one technical property. Privacy and disclosure form the largest individual segment but constitute only slightly more than one quarter of the simulated total. The remaining 74% is distributed across representativeness, provenance, downstream use, purpose alignment, and reproducibility. The central analytical implication is that a privacy-preserving generation mechanism cannot independently establish responsible synthetic-data governance.

This result supports a broader socio-technical interpretation. A dataset may successfully resist a known membership-inference attack and still be unsuitable for collaborative research if it systematically distorts a minority population, lacks version documentation, or is used for questions beyond its validated utility. Conversely, a dataset may exhibit excellent analytical fidelity yet remain unsuitable for wide release because it reproduces rare combinations from the source data too closely. Stadler and colleagues' privacy evaluation and Snoke and colleagues' utility framework demonstrate why these dimensions must be assessed separately.

5.3 Proposed Governance Architecture

The analysis supports a tiered governance architecture rather than a binary distinction between "restricted real data" and "open synthetic data." At the first tier, low-fidelity synthetic datasets can support software development, researcher onboarding, code testing, data-structure familiarization, and educational activities. At the second tier, analytically richer synthetic data can support exploratory modelling under controlled collaborative conditions. At the third tier, high-fidelity synthetic data intended for substantive research require substantially stronger disclosure testing, documentation, and validation. At the final stage, claims intended for publication or policy use should be validated against the protected source dataset whenever the synthetic dataset has not been independently demonstrated to support the exact inferential task. This purpose-sensitive model aligns with ONS guidance that synthetic data should be evaluated according to the intended use rather than through a universal standard of similarity.

A collaborative project should therefore begin with a declared research purpose rather than with the assumption that synthetic data should simply be released as widely as possible. The consortium should identify which variables, relationships, and population structures require preservation, determine acceptable disclosure risk, select an appropriate synthesis technique, document model and software versions, evaluate privacy and utility, establish permitted uses, and specify whether final analysis requires protected-data validation.

6. Discussion

6.1 Synthetic Data Should Be Governed as Derived Research Infrastructure

Synthetic data are sometimes described as though they were merely transformed copies of existing databases. For collaborative research, a more appropriate interpretation is that they constitute derived research infrastructure. Their construction embeds assumptions about what information deserves to be preserved, what information may be distorted, which populations need accurate representation, and how much residual disclosure risk is acceptable. A generative model is therefore not neutral plumbing between a protected database and an external researcher. It is part of the epistemic architecture through which researchers encounter the source population. Snoke and colleagues' distinction between general and specific utility makes this especially clear because the same synthetic dataset can support one analytical purpose while failing another.

Treating synthetic data as infrastructure has important accountability implications. Data producers should document the intended analytical scope, synthesis model, privacy protections, subgroup performance, known weaknesses, software environment, generation date, and version.

Collaborative researchers should not assume that plausible-looking observations imply valid social relationships. They should instead use the documentation to determine whether the dataset was designed to preserve the characteristics their research question requires. The growing concern about inconsistent synthetic-data terminology reinforces the value of standardized documentation because terms such as fidelity, utility, anonymity, and privacy-preserving can otherwise communicate different expectations to different collaborators.

6.2 Privacy Protection and Analytical Fidelity Must Be Evaluated Jointly

The central technical tension in synthetic-data governance concerns the relationship between confidentiality and fidelity. A generator that reproduces the original data extremely closely may produce excellent analytical utility while retaining patterns that enable sensitive inference. A generator that heavily perturbs the source distribution may substantially reduce disclosure risk but also erase correlations, rare groups, interaction structures, and distributional features required for meaningful research. Stadler, Oprisanu, and Troncoso demonstrate empirically that this privacy–utility relationship can be difficult to predict across generative methods, which is precisely why claims of automatic anonymisation are inappropriate.

This trade-off becomes particularly consequential in social research because rare observations are not always statistical noise. Small groups can represent politically, socially, or ethically important populations. Excessively aggressive synthesis may hide precisely those inequalities that researchers seek to understand, while highly faithful synthesis may increase disclosure risk for members of those groups. Governance committees should therefore require subgroup-sensitive utility and risk analysis rather than relying on overall averages.

6.3 Collaborative Access Requires Transparency, Validation, and Continuing Review

Synthetic data can improve collaboration by allowing researchers to begin analytical work while access to protected source data remains restricted, enabling teams to develop code, harmonize models, test workflows, and assess whether a dataset may be suitable for a project. The ONS explicitly recognizes research and testing as valid synthetic-data applications where real data are inaccessible. Yet this benefit depends on transparent communication about what the synthetic version can and cannot establish.

A particularly useful collaborative model is staged validation. Researchers first develop analytical scripts and preliminary models using a synthetic dataset. Once the analysis is stabilized, approved code can be executed against the protected source data in a trusted environment. Differences between synthetic and source results can then be assessed before publication. Snoke and colleagues describe a related “gold standard” approach in which synthetic data support exploratory analysis while final results are obtained from original records under controlled conditions. Such a workflow can reduce unnecessary exposure to sensitive records without treating synthetic findings as automatically definitive.

7. Conclusion

Synthetic data offer a significant opportunity to redesign access to sensitive social research data. By generating artificial records that preserve selected properties of protected datasets, they can enable exploratory analysis, software development, researcher training, methodological collaboration, and

cross-institutional planning without requiring every collaborator to receive immediate access to confidential microdata. Official statistical and regulatory guidance increasingly recognizes these benefits, particularly when real data are difficult to share. At the same time, the same guidance makes clear that synthetic data should not be assumed to reproduce every property of real data or to be automatically anonymous.

The central argument of this study is that synthetic-data governance must integrate privacy and scientific validity. Confidentiality protection alone cannot establish responsible use if the generated data systematically distort population characteristics or produce unreliable research conclusions. Likewise, high analytical fidelity cannot justify unrestricted sharing if sensitive source information remains inferable. The simulated governance analysis reflects this multidimensionality: privacy and disclosure were assigned the largest illustrative contribution, but representativeness, provenance, secondary use, purpose alignment, and reproducibility collectively constituted the majority of overall governance exposure. This pattern reinforces the need to evaluate synthetic data as socio-technical research infrastructure rather than as a single privacy technology.

Collaborative research institutions should consequently implement purpose-sensitive access models. Low-fidelity synthetic data may be adequate for education and programming, while richer datasets intended for substantive analysis require stronger privacy testing, subgroup utility assessment, documentation, and controlled validation. Where researchers use synthetic records to generate publishable claims, the strongest practice is to confirm important findings against the protected source data whenever feasible or clearly disclose the limits of synthetic-only inference. Dataset documentation should state how the data were generated, which properties were evaluated, the intended use, known limitations, release version, and residual disclosure considerations.

Synthetic data therefore should not be presented as a replacement for governance. They are most valuable when they become one component of governance. A well-designed synthetic-data program can reduce unnecessary exposure to confidential records, broaden legitimate collaboration, and support reproducible analytical development, but only when technical generation methods are combined with institutional responsibility, disclosure assessment, utility validation, research ethics, transparent documentation, and continuing review. Future empirical studies should evaluate the proposed framework across multidisciplinary research consortia, compare different synthesis methods across social-science use cases, examine subgroup-level validity, investigate researcher understanding of synthetic-data limitations, and test whether staged synthetic-to-protected-data workflows improve both access efficiency and research integrity. The long-term success of synthetic data in collaborative social science will depend not on how convincingly artificial records resemble real people, but on whether research institutions can demonstrate that those records are sufficiently safe, sufficiently useful, transparently governed, and scientifically appropriate for the questions they are being asked to answer.

References

1. Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468.
2. Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*(3), 441–462.

3. Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Springer.
4. Domingo-Ferrer, J., & Torra, V. (2001). Disclosure risk assessment in statistical data protection. *Journal of Official Statistics, 17*(4), 463–484.
5. Domingo-Ferrer, J., & Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering, 14*(1), 189–201. https://doi.org/10.1109/69.979617
6. Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics, 9*(2), 407–426.
7. Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review, 109*(1), 171–202. https://doi.org/10.1257/aer.20170627
8. Dwork, C. (2006). Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming* (pp. 1–12). Springer. https://doi.org/10.1007/11787006_1
9. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference* (pp. 265–284). Springer.
10. McSherry, F. D., & Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science* (pp. 94–103). IEEE. https://doi.org/10.1109/FOCS.2007.41
11. Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy* (pp. 111–125). IEEE. https://doi.org/10.1109/SP.2008.33
12. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557–570. https://doi.org/10.1142/S0218488502001648
13. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data, 1*(1), Article 3. https://doi.org/10.1145/1217299.1217302
14. Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 106–115). IEEE.
15. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953
16. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems, 32*, 7335–7345.
17. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment, 11*(10), 1071–1083. https://doi.org/10.14778/3231751.3231757
18. Choi, E., Biswal, S., Usry, M., Stewart, W. F., & Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. *Proceedings of Machine Learning for Healthcare, 68*, 286–305.