Data Privacy in the AI Era and Implications for Survey Practice[1]

Shurong Lin[1] **and Aleksandra Slavković**[2]

Department of Statistics, The Pennsylvania State University, USA
[1]shurong@psu.edu, [2]sesa@psu.edu

Abstract

Balancing data confidentiality and data utility is a long-standing challenge in survey and official statistics. In modern data ecosystems, this challenge has intensified due to increased data availability, advances in computational methods, and now the widespread use of artificial intelligence (AI). The rise of AI does not create the familiar privacy–utility tradeoff anew, but fundamentally reshapes the information environment in which it must be understood. Protected survey data can now be reused, linked, and modeled more intensively than before, while AI-assisted tools are transforming survey data collection, processing, and analysis. In this paper, we present a perspective grounded in statistical data privacy (SDP), an integrative framework that connects traditional statistical disclosure control with formal privacy approaches such as differential privacy. We emphasize that privacy-preserving mechanisms alter released data or summaries by introducing additional sources of uncertainty that must be explicitly modeled to support valid downstream analysis and statistical inference. This perspective is important in the AI era, especially for surveys where complex design and processing features already shape inference. AI further reshapes the privacy–utility tradeoff by enabling adaptive reuse and accelerating the accumulation of privacy risk. This reinforces the need for methodological transparency and alignment of privacy choices with the policy goals and intended uses of survey data.

Keywords: Artificial intelligence (AI), statistical data privacy, differential privacy, statistical disclosure control, survey statistics, official statistics.

1 Introduction

Statistical agencies, survey organizations, and data stewards have long operated under a dual mandate: to protect the confidentiality of respondents while ensuring that released data remain sufficiently informative to support statistical inference, scientific discovery, and policy. This balance is foundational to the mission of official statistics and has guided decades of methodological development (e.g., see Hundepool et al., 2012; Skinner, 2009; Matthews and Harel, 2011). The tension between minimizing disclosure risk and maximizing data utility has traditionally been framed as a *risk–utility tradeoff* (Duncan, Keller-McNulty, and Stokes, 2001). In this article, we adopt the broader term *privacy–utility tradeoff* and use the terms *privacy* and *confidentiality* interchangeably to refer to protection against the disclosure of sensitive information by a third party.[2]

The privacy-utility tradeoff is especially difficult to attain in survey settings. Survey products are often intended to support finite-population inference under complex designs, unequal weighting, nonresponse adjustment, imputation, and other processing steps (Groves, F. J. Fowler, et al., 2011; Särndal, Swensson, and Wretman, 1992; Lohr, 2021). For decades, the survey community has worked

[1]This paper draws on the 2025 Links Lecture by Slavković: *PrAIvacy (noun) Pronounced: /'prā-və-sē/ Balancing data confidentiality and utility* (https://www.amstatleads.org/events).

[2]Informally, privacy is the right of individuals to control the dissemination of, or access to, information about themselves. Confidentiality is the dissemination of data without public identification (i.e., no disclosure) by a third party.

with a broad collection of methods from statistical disclosure control (SDC), also referred to as statistical disclosure limitation (Hundepool et al., 2012; Matthews and Harel, 2011). More recently, formal privacy frameworks such as differential privacy (DP) (Dwork, McSherry, et al., 2006; Dwork and Roth, 2014) have received attention. A major example in official statistics is the adoption of DP for the 2020 U.S. Decennial Census (J. M. Abowd, 2018; J. Abowd et al., 2022). We do not review the formal definition of DP here; for a recent overview in this venue, see Charest and Drechsler, 2026. We use the term *statistical data privacy* (SDP) to refer to methods and frameworks for protecting confidential information in data release and analysis, including both SDC and DP (Slavković and Seeman, 2023).

Traditionally, sample surveys were designed to produce tabular estimates from a single dataset. Early total survey error frameworks focused primarily on measurement and representation errors and did not explicitly address non-sampling errors induced by SDC (Groves, J. Fowler F. J., et al., 2004). Later survey quality frameworks acknowledged that disclosure-control procedures can affect data quality, but did not fully address their implications for valid statistical inference (Groves and Lyberg, 2010). Over the past two decades, however, survey practice has shifted toward statistically valid estimation based on the integration of multiple data sources, with such combined data products becoming increasingly common. At the same time, the heightened disclosure risks inherent in these integrated settings have not received comparable attention, nor has the role of privacy-induced errors in shaping valid statistical inference.

In today's data ecosystem, maintaining the privacy–utility balance has become increasingly difficult. Data are more abundant, interconnected, and reusable than ever before. Advances in computational power and the availability of auxiliary data have made it possible to combine information across sources, often in ways not anticipated at the time of collection. Early work showed that even aggregate data releases could enable the reconstruction of sensitive information (e.g., Dinur and Nissim, 2003), and more recent work has demonstrated that machine learning models can leak training data or enable inference attacks (e.g., Shokri et al., 2017; Carlini et al., 2021).

Artificial intelligence (AI) has amplified both the opportunities and the risks associated with data. Here, AI refers broadly to machine-learning systems that automate, augment, or support analytical and decision-making tasks, including generative models, large language models, and tools for prediction, classification, linkage, and data generation. On one hand, AI enables richer analysis, improved prediction, and new forms of data integration. On the other hand, it challenges traditional assumptions about how data are used, reused, and shared. AI changes the information environment in which the privacy-utility tradeoff must be understood. Privacy can no longer be viewed as a static property of a dataset; instead, it must be understood as a dynamic feature of a broader data ecosystem.

In survey settings, AI-driven systems can affect multiple stages of the survey pipeline, from design and respondent interaction to processing, analysis, reporting, and synthetic data generation (Rothschild et al., 2025). They can also search across heterogeneous sources and exploit auxiliary information at scale. Released data products, including synthetic data, can therefore be reused, linked, and modeled far more intensively than before (Kapania et al., 2025). As a result, privacy depends not only on whether a released product satisfies a disclosure rule at the time of release, but also on how it is used in downstream analyses and combined with external data. Privacy risk may therefore accumulate and evolve over time. In the AI era, overlooking amplified disclosure risks in combined-data settings becomes especially consequential.

This paper approaches these challenges through the lens of statistical data privacy (SDP), an integrative framework that connects traditional SDC with formal privacy approaches such as DP. SDP is particularly well suited to survey statistics because it emphasizes uncertainty quantification, inferen-

tial validity, and transparency. From a statistical perspective, privacy protection methodology should consider both the disclosure-control aspect and the quality of inference (J. M. Abowd and Schmutte, 2015; Slavković and Seeman, 2023; Awan and Gong, 2024). Privacy-preserving methods and mechanisms, whether based on SDC or DP frameworks, typically alter released data or summaries by introducing additional sources of uncertainty that must be explicitly modeled to support valid downstream analyses. Once confidentiality protection alters released data, summaries, or estimators, standard inferential procedures based on the protected data no longer automatically retain their validity. Privacy protection thus becomes an integral component of the broader error and uncertainty structure that shapes survey inference.

The SDP perspective is valuable for several reasons. First, it provides a common statistical language for understanding both traditional disclosure limitation and modern formal privacy methods. Second, it aligns naturally with the survey community's longstanding focus on multiple sources of error, including sampling variation, nonresponse, weighting, imputation, and linkage. Third, it highlights how AI increases the scale of data reuse and the adaptivity of downstream analysis, making the inferential consequences of privacy protection more important, not less.

In the remainder of the paper, we outline how the challenges we face today are rooted in a long history of statistical practice. We propose that the survey community should treat privacy protection as an essential component of the statistical pipeline through which data are produced, released, and analyzed. In the AI era, this perspective allows us not only to revisit the classic privacy–utility tension, but also to ask a more informative question: under what forms of protection can released data support trustworthy analysis, for which inferential goals, and with what accounting of uncertainty?

2 Privacy in the Era of AI

2.1 A Bit of a Historical Reflection

In thinking about how best to quantify the data privacy–utility tradeoff, we can view AI as an amplifier that stress-tests prior paradigms. The challenges we face today are rooted in a long history of statistical practice. The timeline in Table 1 highlights more than 70 years of intellectual and practical development, particularly within the U.S. federal statistical system, as it has responded to growing data demands. It also illustrates how AI exposes and challenges the assumptions and protections associated with each stage.

Historically, privacy benefited from structural constraints. Data demand was relatively limited, data systems were slow and fragmented, and computation was expensive. Early statistical confidentiality practices often relied implicitly on these constraints for protection. Prior to the 1960s, agencies developed tabulation rules and heuristics to prevent disclosure. These approaches were effective in an environment with limited access and restricted analytical capabilities.

The 1960s marked a turning point with advances in computing, which enabled the development of more systematic statistical disclosure control (SDC) methods. Although data access remained restricted and large-scale linkage across sources was still difficult in practice, increasing computational capacity began to expose vulnerabilities in earlier approaches. Methods such as data swapping and complementary cell suppression addressed the needs of the time, but their outputs became more susceptible to linkage attacks (e.g., Sweeney, 2002) and database reconstruction (Dinur and Nissim, 2003) as computational capabilities improved. Holan et al., 2010 and Webb et al., 2026, among others, also demonstrate strong deficiencies in privacy protection with cell suppression.

As data sharing expanded and computational resources grew, SDC methods evolved to address a setting in which more granular data were requested, auxiliary information became more widely available, and record linkage and inferential reconstruction became increasingly feasible (e.g., Homer et al., 2008; A. Slavkovic and Lee, 2010; Hundepool et al., 2012; Gymrek et al., 2013). The emergence of model-based approaches marked an important shift. Synthetic data and multiple imputation introduced the idea that confidentiality protection could be integrated with statistical modeling (Rubin, 1993; Reiter, 2005), aiming to preserve key statistical properties while reducing disclosure risk. These approaches were later extended to large-scale applications (Kinney et al., 2011).

With the digital age, the growth of data and the increasing availability of diverse forms of public data beyond agency releases increased re-identification risks (Narayanan and Shmatikov, 2008; Dwork, Smith, et al., 2017). Differential privacy represents a subsequent evolution of statistical data privacy, providing a formal framework for bounding disclosure risk through controlled randomness (Dwork, McSherry, et al., 2006). It builds on earlier ideas such as noise injection, while introducing *transparency, composability, and formal guarantees* (Dwork and Roth, 2014). Over the past two decades, differential privacy has become an influential framework, in part because intuitive or ad hoc disclosure-control rules appeared increasingly fragile in data-rich environments with abundant side information and powerful computation. It has also seen large-scale implementation in official statistics, including at the U.S. Census Bureau (J. M. Abowd, Ashmead, et al., 2022) as well as in industry settings such as Apple and Google (Apple, 2017; Erlingsson, Pihur, and Korolova, 2014). At the same time, researchers have been working on the broader challenge of principled integration of DP and other formal privacy guarantees with decades of SDC experience in data utility and statistical approaches to uncertainty quantification.

Today, the emergence of AI represents another shift. AI enables learning, data generation, and reuse at an unprecedented scale. The structural constraints that once limited data reuse have largely disappeared, and privacy risk must now be understood as evolving over time. In the current AI era, the central question is no longer only whether a single data release is safe at the moment of dissemination. Instead, released data may enter broader workflows in which they are reused, transformed, linked with auxiliary sources, and incorporated into downstream models and generated outputs. As a result, privacy risk becomes increasingly temporal, adaptive, and cumulative. At the same time, while these developments expand what can be learned from data, the errors and biases introduced by both learning algorithms and privacy-preserving mechanisms may also accumulate. This raises further questions about data utility, including accuracy, precision, credibility, and relevance.

Table 1: How AI stress-tests long-standing privacy challenges across successive eras of confidentiality protection.

Era	What changed	How AI stress-tests it
Early confidentiality (pre-1960s)	Low data reuse	AI enables large-scale inference from aggregates
Traditional SDC (1960s–1990s)	Limited linkage	AI automates linkage and reconstruction
Modern SDC and synthetic data (early 2000s)	Valid inference	AI increases model capacity and can raise re-identification risk
Differential privacy (mid-2000s)	Worst-case guarantees	AI makes worst-case attackers more realistic
AI (now)	**Learning everywhere**	**Privacy risk is temporal, adaptive, and cumulative**

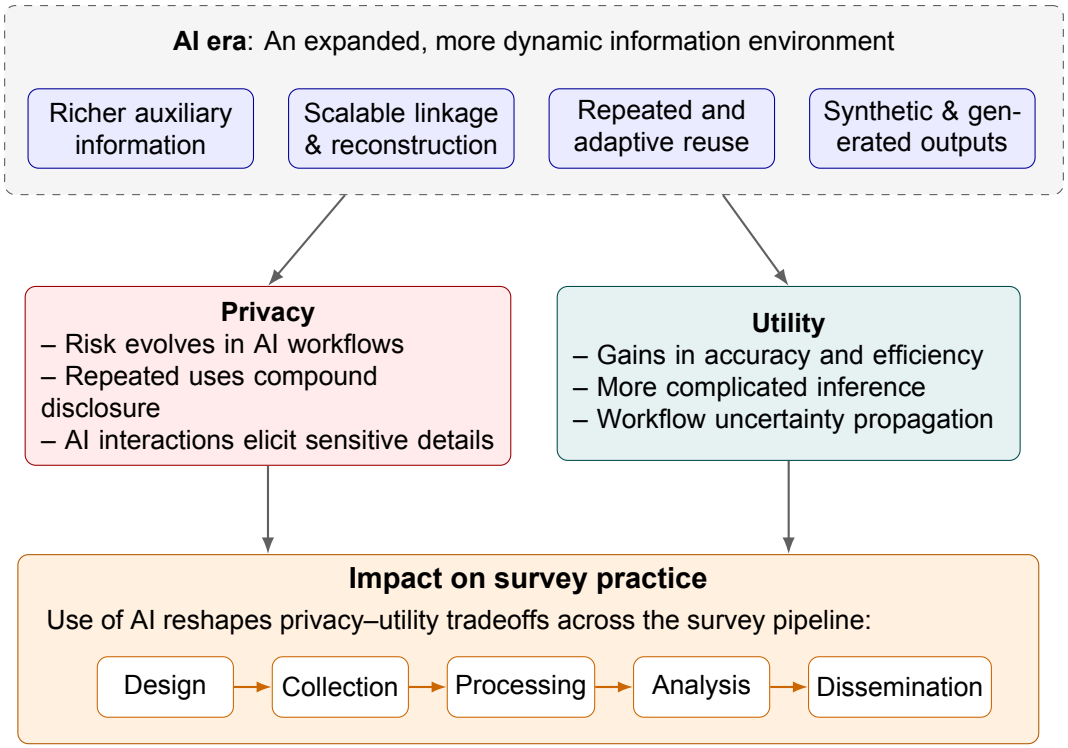

Figure 1: How the AI era reshapes privacy and utility and their implications across the survey pipeline.

2.2 AI as an Amplifier of Both Utility and Risk

AI raises the stakes of data privacy. In earlier settings, privacy protection was often framed as a release-stage problem, in which a table, statistic, or synthetic dataset was produced, evaluated, and then disseminated. In the AI era, this view is too narrow. Released data may instead enter broader information environments, where they can be queried repeatedly, linked to other sources, incorporated into model-based systems, and used to generate new outputs. This expanded setting makes robust and transparent privacy protection more critical, particularly because informal disclosure-control rules may become fragile when data are reused and combined in unforeseen ways. While this shift is not unique to survey data, it is especially consequential in survey settings, where these dynamics affect multiple stages of the survey pipeline, including questionnaire design, data collection, processing, analysis, and dissemination. Figure 1 summarizes, at a high level, these implications for the privacy–utility tradeoff.

On the privacy side, released survey products may now be combined with richer auxiliary information, linked or reconstructed at scale, reused repeatedly and adaptively, or transformed into synthetic and other generated outputs. As a result, privacy risk becomes increasingly dependent on the surrounding information environment rather than solely on the released object itself. It may evolve as data move through AI-enabled workflows, with repeated use compounding disclosure concerns. In conversational or chatbot-based survey administration, AI may also elicit richer and potentially more sensitive respondent information than a fixed questionnaire would have collected (Wuttke et al., 2025; Xiao et al., 2020).

On the utility side, AI-assisted tools can improve efficiency and, in some settings, task-level accuracy by supporting question development, coding open-ended responses, summarizing complex outputs, and making survey results more accessible to broader audiences (Törnberg, 2023; Mellon et al., 2024). These tools may also reduce costs by automating or augmenting labor-intensive steps in the

survey workflow (Jansen, Jung, and Salminen, 2023). In addition, synthetic data and other model-generated outputs are increasingly discussed as ways to expand access while reducing direct exposure of confidential data (Kapania et al., 2025). Such products may support exploration, software development, education, early-stage analysis, or model prototyping.

These potential gains in utility do not automatically translate into inferential validity. In survey settings, utility depends not only on whether specific steps become faster or more accurate, but also on whether the resulting data products support credible estimation and appropriate uncertainty quantification. AI-enabled workflows can propagate errors across stages: inaccuracies in coding may affect downstream estimates, synthetic outputs may preserve certain distributional features while distorting inferential targets, and automated reporting tools may present results without clearly conveying uncertainty arising from sampling, processing, and privacy protection. As a result, AI may expand the practical uses of protected survey data while making the validity of inference more difficult to assess.

These issues are especially important for survey data because such data occupy a distinctive role in the broader data ecosystem. Survey data are carefully curated, closely tied to population-level inference and public policy, and often contain sensitive information. They are also produced through a complex inferential pipeline shaped by sampling, weighting, editing, imputation, calibration, and, in some cases, record linkage. In this context, privacy protection cannot be treated as a simple technical add-on; it interacts directly with data production, downstream use, and statistical interpretation.

Recent discussions in the survey community have highlighted both the opportunities and the risks of AI. While AI may enhance many aspects of survey work, it also raises concerns about transparency, validation, and the reliability of generated outputs (Rothschild et al., 2025). These concerns align naturally with privacy. As it becomes increasingly difficult to distinguish between original data, protected releases, and generated outputs, it becomes more important to understand how privacy protection enters the survey pipeline and how it affects both disclosure risk and inferential utility.

3 Privacy, Uncertainty, and Inference

Statistical inference typically reflects a sequence of data-generation, collection, and processing steps rather than a single-step calculation. Privacy protection fits naturally into this broader framework because it systematically modifies the released object on which downstream analysis is based; see, for example, Figure 1 in Slavković and Seeman (2023). In this section we outline a perspective that the additional uncertainty introduced by privacy protection is both unavoidable and essential to account for in order to support valid statistical inference. This perspective extends beyond survey data alone, but it is especially important in survey settings, where inference is already shaped by multiple sources of uncertainty, including sampling variation, nonresponse, weighting, imputation, and linkage (Groves, F. J. Fowler, et al., 2011).

3.1 Privacy Risk and Statistical Utility

A useful way to formalize privacy and inference is to distinguish among the confidential data object, the released (sanitized) object, and the information available to different data users. Here, *privacy (disclosure) risk* refers to what can be learned about individuals or confidential records from a release, whereas *statistical utility* refers to what the release still supports in terms of estimation, uncertainty quantification, and downstream analysis.[3]

[3] Our notation aligns with Seeman (2023).

Let $\theta \in \Theta$ denote the population quantity or model parameter of interest, $f_\theta(\cdot)$ the data-generating model, $X \in \mathcal{X}$ the confidential data, and $Y \in \mathcal{Y}$ the released output after privacy protection. We write $M_X(\cdot)$ for the privacy-preserving release mechanism or method. In DP, such mechanisms are typically randomized, for example through noise addition; in SDC, they may be randomized, as in data swapping or noise addition, or deterministic, as in recoding. Then we write

$$X \sim f_\theta, \qquad Y = M_X(X). \tag{1}$$

Let $\pi_A(\cdot)$ and $\pi_D(\cdot)$ denote, respectively, the information available to an adversary and to a data analyst. Here, an adversary refers to a party attempting to infer sensitive information about individuals or confidential records from released data and possible auxiliary information. Privacy risk can then be viewed through the adversary's updated information,

$$\pi_A(\cdot \mid M_X, Y), \tag{2}$$

whereas statistical utility may be viewed through the analyst's resulting estimator,

$$\hat{\theta} = \hat{\theta}(M_X, Y, \pi_D). \tag{3}$$

This framework naturally accommodates additional information sources. Let Z denote external data, such as linked datasets, auxiliary information, prior or related releases, or AI-generated inputs. Such information may improve downstream analysis, but in both SDC and DP settings it can also increase privacy risk by facilitating record linkage, attribute inference, or reconstruction across releases.

The cumulative perspective is particularly challenging for SDC, where protection is often assessed for a specific release under context-dependent assumptions about what an intruder may know. As the surrounding information environment evolves, these assumptions may become unstable. As discussed in Slavković and Seeman (2023), SDC typically operates with an *absolute* notion of disclosure risk, whereas DP defines a *relative* notion of risk over a broader set of adversarial contexts, quantifying differences in disclosure risk between similar datasets. In the AI era, privacy risk becomes increasingly *relative* to the broader information environment, as it depends on prior releases, auxiliary data, and patterns of downstream reuse.

Differential privacy provides a clear illustration of this point. Relaxed notions such as (ϵ, δ)-DP permit a small failure probability, and repeated releases affect not only the cumulative privacy-loss parameter ϵ but also the failure-probability term δ. Thus, while DP provides a principled framework for accounting for repeated analyses, the cumulative risk may still grow over time. More broadly, privacy concerns may increase in mixed release systems that combine protected and unprotected outputs. For example, an earlier release may provide unprotected state-level summaries, while a later release provides protected county-level outputs derived from the same underlying data. Although the latter release may satisfy its formal privacy guarantee in isolation, disclosure risk may increase when the two are interpreted jointly. This motivates formal approaches to partially private data (Seeman, Reimherr, and Slavković, 2022). Taken together, these considerations suggest that privacy should be evaluated not only at the level of individual outputs, but also at the level of the broader system of linked data products, related releases, and external information.

3.2 Two Inferential Regimes

There are two inferential regimes for analysis under privacy protection. In Slavković and Seeman (2023), these are formulated as statistical risk minimization problems, where $L(\cdot, \cdot)$ denotes a loss

function. In what they call the *design problem*, the privacy mechanism and the estimator are chosen jointly for a particular inferential goal. A generic formulation is

$$\hat{\theta}_{\text{Design}} = \arg \min_{\tilde{\theta},\, M \in \mathcal{M}} \sup_{\theta \in \Theta} E_{\theta,M} \left[L\{\tilde{\theta}(Y), \theta\} \right], \tag{4}$$

where Θ denotes the parameter space and $\mathcal{M}$ denotes a class of admissible privacy-preserving mechanisms. For a fixed data-generating parameter θ and mechanism $M \in \mathcal{M}$, $E_{\theta,M}$ denotes expectation with respect to both the data-generating process indexed by θ and any randomness introduced by the mechanism M.

By contrast, in the *adjustment problem*, the privacy mechanism is fixed in advance and the task is to construct valid inference from the released object. In that case,

$$\hat{\theta}_{\text{Adjust}} = \arg \min_{\tilde{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta,M} \left[L\big(\tilde{\theta}(Y), \theta\big) \right]. \tag{5}$$

These formulations clarify the distinction at a formal decision-theoretic level and serve as useful conceptual targets. In practice, however, privacy-preserving methods are rarely obtained by explicitly solving these minimization problems, which are often intractable in realistic settings. Instead, methods are typically developed through problem-specific constructions. For example, a wide range of methods has been proposed for differentially private linear regression, and their performance varies across data settings and assumptions rather than yielding a single uniformly optimal solution (Wang, 2018; Amin et al., 2023; Lin, Slavković, and Bhoomireddy, 2026). For this reason, it is often more useful to describe the corresponding settings in terms of workflow. We refer to these as *primary analysis under privacy* and *secondary analysis under privacy*. This terminology emphasizes the practical distinction most relevant for analysts: whether privacy protection is incorporated into the statistical procedure from the outset or instead enters through an already protected release that must be analyzed afterward. A closely related distinction appears in record linkage, where one differentiates between settings in which linkage and analysis are handled jointly and settings in which analysts are provided with linked data for downstream analysis (Kamat and Gutman, 2026). The same lesson carries over to privacy: uncertainty introduced during data construction may either be incorporated directly into the analysis or addressed only after the fact.

Under privacy protection, *primary analysis* refers to settings in which the release mechanism, inferential target, and uncertainty quantification are developed jointly, whereas *secondary analysis* refers to settings in which the analyst is given a protected release and must determine what valid inference remains possible. The distinction is therefore not only chronological but also concerns what is under the analyst's control. This makes transparency and documentation especially important, since reliable inference depends on understanding how the protected object was constructed and what perturbations it reflects.

3.3 Inferential Error and Uncertainty

Privacy protection matters for inference because it can affect both variability and bias. These quantities are especially important in survey settings, where they directly shape the quality of point estimates, interval estimates, and other inferential summaries used in policy and official statistics. Once Y is generated from the confidential data through a randomized release mechanism, its variability reflects both the underlying data-generating and sampling process and the additional randomness

introduced by the release mechanism. For randomized mechanisms, a useful decomposition is

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\mathbb{E}[Y \mid X]\right) + \mathbb{E}\left[\mathrm{Var}(Y \mid X)\right], \tag{6}$$

where the outer expectation and variance are taken with respect to the data-generating and sampling process for X, while the conditional expectation and variance are taken with respect to the randomness of the privacy mechanism given X. The first term captures variability inherited from the data-generating and sampling process, while the second reflects variability introduced by privacy protection.

To isolate the effect of privacy protection on bias, it is useful to compare the release Y to the corresponding non-private quantity $T(X)$ for the same estimation goal. The privacy-induced bias conditional on X may then be written as

$$b_{\mathrm{priv}}(X) = \mathbb{E}[Y \mid X] - T(X). \tag{7}$$

Averaging over the data-generating process gives

$$\mathbb{E}[Y] - \theta = \mathbb{E}\left[\mathbb{E}[Y \mid X] - T(X)\right] + \left(\mathbb{E}[T(X)] - \theta\right), \tag{8}$$

where the first term represents bias introduced by the privacy mechanism and the second corresponds to the bias of the non-private estimator. Valid inference therefore requires both sources of error to be accounted for.

These considerations are particularly evident in confidence interval (CI) estimation. If privacy-induced noise is ignored and the released statistic is analyzed as though it were the corresponding non-private statistic, the resulting inference will likely be adversely affected in ways that may be difficult for the analyst or data user to detect. Confidence intervals may be too narrow, empirical coverage may fall below the nominal level, and point estimates may exhibit bias arising from naive downstream analysis. These issues extend beyond confidence intervals to more general inferential procedures.

Figure 2 illustrates this phenomenon in a linear regression setting using synthetic data generated by BinAgg (Lin, Slavković, and Bhoomireddy, 2026), a differentially private synthetic data method paired with a customized regression procedure. The first panel shows the sampling distribution of the non-private OLS estimator for the regression coefficient and its corresponding CI. The second and third panels show the distributions for estimators based on differentially private synthetic data. The second panel corresponds to a naive approach that fits linear regression directly to the synthetic data, treating them as if they were the original data. The third panel corresponds to the BinAgg approach, which adjusts regression to account for the additional variability introduced by differential privacy.

Both private approaches exhibit greater dispersion than the non-private OLS benchmark, since the synthetic data reflect not only sampling variability but also additional variability induced by the privacy mechanism. The naive analysis of the synthetic data exhibits both noticeable bias and undercoverage of its CIs, whereas the BinAgg estimator is much closer to unbiasedness and its adjusted CI attains coverage close to the nominal level. In fact, the associated large sample theory based on a central limit theorem is designed to justify asymptotic unbiasedness and valid interval estimation under the BinAgg procedure.

The AI-era setting described in Section 2 makes careful inferential accounting even more important. Protected releases are increasingly reused in automated workflows, combined with external information, and incorporated into model-based systems. As a result, the uncertainty introduced by privacy

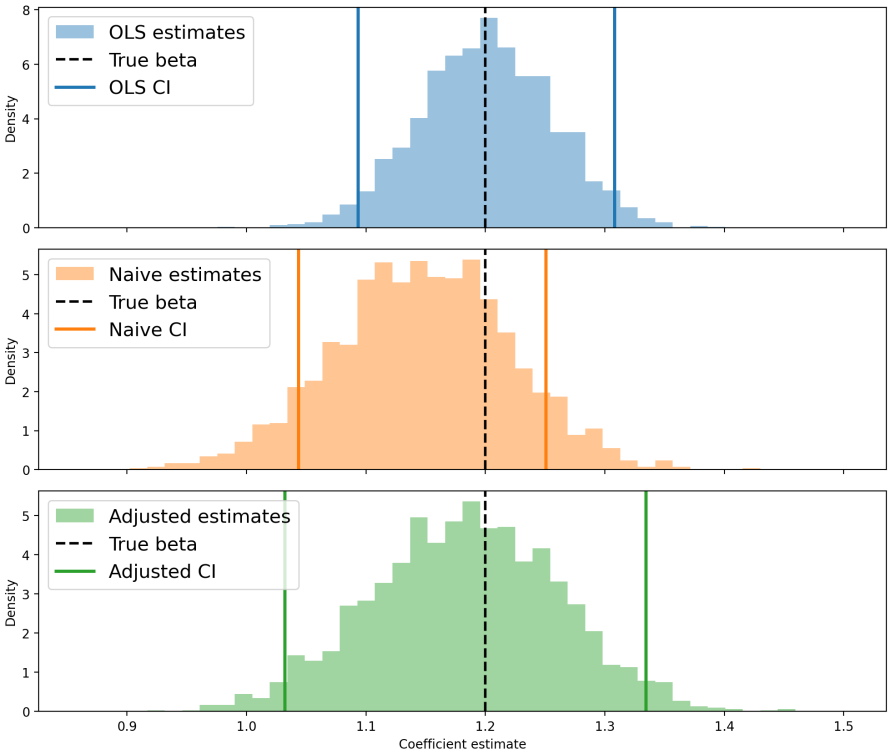

Figure 2: Confidence intervals for the regression coefficient under non-private and differentially private synthetic-data-based simple linear regression for a sample of size $n = 1000$. Synthetic data generated by BinAgg under μ-Gaussian DP (Dong, Roth, and Su, 2022) with $\mu = 1$.

protection may propagate through subsequent stages of processing and analysis rather than remaining confined to the initial release. The inferential consequences of privacy protection therefore extend beyond the point of release and carry into downstream analysis.

4 Statistical Data Privacy Implications for Survey Practice

Viewing privacy protection as part of the data pipeline has important implications for survey practice. Survey data are produced through a sequence of design and processing decisions, including questionnaire design, sample design, field procedures, weighting, calibration, editing, imputation, and, in some cases, record linkage. Each of these steps helps determine what can ultimately be learned from the data.

These design and processing choices are typically aligned with the intended end uses of the survey data. For example, a survey designed primarily to publish tabulations or change estimates may place greater emphasis on preserving valid totals and comparisons, whereas a survey designed to release research microdata may place greater emphasis on preserving broader distributional features. These priorities, in turn, affect both disclosure risk and the choice of privacy protection methods. In this sense, privacy protection is not external to the survey process, but part of the same chain of design and processing decisions. Because privacy protection modifies the released data product, it affects not only disclosure risk, but also the inferential validity, interpretability, and usability of the resulting data products.

It is therefore important to understand how privacy mechanisms interact with survey-specific structures. Complex sample designs, including stratified, clustered, multistage, and unequal-probability

designs, can affect both the sensitivity and the variability of survey estimators. Survey weights may amplify the influence of certain units and change how privacy-induced perturbations propagate through estimation. Nonresponse adjustment, calibration, editing, and imputation introduce additional dependencies and processing layers. Together, these features complicate the design of privacy mechanisms and make it more difficult to determine how privacy-induced perturbations interact with downstream estimation. Existing work has begun to address how specific survey features interact with differential privacy, including imputation, stratified sampling, survey weighting, and record linkage (Das et al., 2022; Lin, Bun, et al., 2024; Seeman, Si, and Reiter, 2026; Lin, Paquette, and Kolaczyk, 2024). For a broader discussion of the challenges of applying differential privacy in survey settings, see Drechsler and Bailie (2026). More generally, privacy must be understood in conjunction with the survey design and processing steps that shape the data object used for analysis.

Another implication is the need for transparency (**Gong2022Transparent**) and documentation, especially for secondary analysis under privacy. Survey methodologists have long emphasized documenting sampling, weighting, nonresponse adjustment, editing, and imputation because these features directly affect inference. Privacy protection should be documented with the same level of care. This does not imply that every operational detail must be disclosed. Full disclosure of SDC procedures is often impractical, whereas DP emphasizes algorithmic transparency. In either case, analysts need sufficient information to understand how privacy protection affects inference. Without such information, protected releases may be treated as if they were unmodified data, leading to misscalibrated downstream analyses and inference; e.g., for simple examples, see A. Slavkovic and Lee, 2010 for the case of two-way contingency table analysis or Woo and A. B. Slavkovic, 2012 for a logistic regression estimation after PRAM as a disclosure avoidance method was applied.

This issue is especially important for synthetic data products. In the generative AI era, such products are increasingly presented as a response to data scarcity, resource constraints, and limits on access to sensitive data (Kapania et al., 2025). Because they are often positioned as accessible alternatives to restricted data, transparency about how they are generated, which features they preserve, and what inferential tasks they are intended to support becomes essential.

A further implication is that privacy choices should be aligned with the policy goals and intended uses of the survey. Survey data are often produced to support population estimates, subgroup comparisons, trend monitoring, and policy decisions. These objectives are not interchangeable, and they are not equally sensitive to privacy protection. A release mechanism that preserves broad national patterns may still distort local heterogeneity, weaken estimation for small areas and subpopulations, or reduce reliability for rare outcomes. In practice, this means that privacy protection cannot be evaluated solely in terms of overall utility or average error. It must be assessed relative to the specific estimands, disaggregations, and inferential tasks the survey is intended to support. In survey settings, privacy design is therefore partly a matter of prioritization: deciding which uses must remain reliable, for whom, and with what accounting of the uncertainty introduced by protection.

5 Conclusion

To unlock the transformative potential of AI, data privacy must remain central to digital strategy. It is fundamental to ethical and socially responsible AI and machine learning. Statistical data privacy methods provide a coherent framework for this setting by combining key properties of formal privacy, such as methodological transparency, robustness to post-processing, and principled accounting of cumulative privacy loss, with careful attention to statistical utility. Privacy-preserving algorithms introduce structured randomness that, if not properly accounted for, can distort downstream analysis, amplify

the effects of data scarcity, and introduce bias. Uncertainty quantification must therefore remain a core principle when working with protected data and when generating curated synthetic datasets. At the same time, AI makes uncertainty quantification more complex and privacy risk more continuous, adaptive, and cumulative.

The AI era amplifies familiar privacy concerns in survey and official statistics by making them more dynamic, more cumulative, and more tightly connected to downstream analysis. Privacy protection must therefore be evaluated not only in terms of disclosure risk, but also in relation to the inferential goals that released data are intended to support. This is especially important in survey settings, where data are produced through multiple stages of design and processing and are often intended to support inference by secondary users. This perspective highlights the importance of survey-specific design features, transparent documentation, and evaluation relative to intended uses.

The survey community brings important conceptual and technical tools to these challenges, including design-based thinking, careful uncertainty quantification, and a sustained focus on intended use. In the AI era, the central question is not simply whether privacy protection reduces risk or utility, but under what forms of protection, for which inferential goals, and with what accounting of uncertainty released data can still support trustworthy, transparent, and policy-relevant analyses in an increasingly complex data ecosystem.

Acknowledgment

This work was supported at Penn State by the Huck Institutes of the Life Sciences through the Dorothy Foehr Huck and J. Lloyd Huck Chair in Data Privacy and Confidentiality, and by a 2025–2026 Rising Researcher Grant from the Institute for Computational and Data Sciences (RRID: $SCR025154$).

References

Abowd, J., R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala, B. Moran, W. Sexton, M. Spence, and P. Zhuravlev (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. In: *Harvard Data Science Review*. Special Issue 2.

Abowd, J. M. (2018). The U.S. Census Bureau Adopts Differential Privacy. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867.

Abowd, J. M., R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala, B. Moran, W. Sexton, M. Spence, and P. Zhuravlev (June 2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. In: *Harvard Data Science Review*. Special Issue 2. URL: https://hdsr.mitpress.mit.edu/pub/7evz361i.

Abowd, J. M. and I. M. Schmutte (2015). Economic Analysis and Statistical Disclosure Limitation. In: *Brookings Papers on Economic Activity* 50.1 (Spring), pp. 221–267. URL: https://www.brookings.edu/about/projects/bpea/papers/2015/economic-analysis-statistical-disclosure-limitation.

Amin, K., M. Joseph, M. Ribero, and S. Vassilvitskii (2023). Easy Differentially Private Linear Regression. In: *International Conference on Learning Representations (ICLR)*.

Apple (2017). *Learning with Privacy at Scale*. URL: https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

Awan, J. and R. Gong (2024). "Statistical Inference and Differential Privacy". In: *Handbook of Sharing Confidential Data*. Chapman and Hall/CRC, pp. 115–135.

Carlini, N., F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel (Aug. 2021). "Extracting Training Data from Large Language Models". In: *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, pp. 2633–2650. ISBN: 978-1-939133-24-3. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Charest, A.-S. and J. Drechsler (2026). Differential Privacy and its Application to Survey Data. In: *The Survey Statistician* 93, pp. 13–25.

Das, S., J. Dreschler, K. Merrill, and S. Merrill (2022). Imputation under Differential Privacy. In: *ArXiv* abs/2206.15063. URL: https://api.semanticscholar.org/CorpusID:250144515.

Dinur, I. and K. Nissim (2003). "Revealing information while preserving privacy". In: *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '03. San Diego, California: Association for Computing Machinery, pp. 202–210. ISBN: 1581136706. DOI: 10.1145/773153.773173. URL: https://doi.org/10.1145/773153.773173.

Dong, J., A. Roth, and W. J. Su (Feb. 2022). Gaussian Differential Privacy. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84.1, pp. 3–37. DOI: 10.1111/rssb.12454.

Drechsler, J. and J. Bailie (2026). "The Complexities of Differential Privacy for Survey Data". In: *Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and New Findings*. University of Chicago Press. Chap. 5. URL: https://www.nber.org/books-and-chapters/data-privacy-protection-and-conduct-applied-research-methods-approaches-and-new-findings/complexities-differential-privacy-survey-data.

Duncan, G. T., S. A. Keller-McNulty, and S. L. Stokes (2001). *Disclosure Risk vs. Data Utility: The R-U Confidentiality Map*. Tech. rep. Technical Report Number 121. Research Triangle Park, NC: National Institute of Statistical Sciences.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In: *Theory of Cryptography Conference*, pp. 265–284.

Dwork, C. and A. Roth (Aug. 2014). The Algorithmic Foundations of Differential Privacy. In: *Foundations and Trends in Theoretical Computer Science* 9.3–4, pp. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042. URL: https://doi.org/10.1561/0400000042.

Dwork, C., A. Smith, T. Steinke, and J. Ullman (2017). Exposed! A Survey of Attacks on Private Data. eng. In: *Annual Review of Statistics and Its Application (2017)*.

Erlingsson, Ú., V. Pihur, and A. Korolova (2014). "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. CCS '14. Scottsdale, Arizona, USA: Association for Computing Machinery, pp. 1054–1067. ISBN: 9781450329576. DOI: 10.1145/2660267.2660348. URL: https://doi.org/10.1145/2660267.2660348.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2011). *Survey Methodology*. 2nd ed. Wiley.

Groves, R. M., J. Fowler Floyd J., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey Methodology*. New York: Wiley.

Groves, R. M. and L. Lyberg (2010). Total Survey Error: Past, Present, and Future. In: *Public Opinion Quarterly* 74.5, pp. 849–879. DOI: 10.1093/poq/nfq065.

Gymrek, M., A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich (2013). Identifying Personal Genomes by Surname Inference. In: *Science* 339.6117, pp. 321–324. DOI: 10.1126/science.1229566. eprint: https://www.science.org/doi/pdf/10.1126/science.1229566. URL: https://www.science.org/doi/abs/10.1126/science.1229566.

Holan, S. H., D. Toth, M. A. R. Ferreira, and A. F. Karr (2010). Bayesian Multiscale Multiple Imputation With Implications for Data Confidentiality. In: *Journal of the American Statistical Association*

105.490, pp. 564–577. DOI: 10.1198/jasa.2009.ap08629. eprint: https://doi.org/10.1198/jasa.2009.ap08629. URL: https://doi.org/10.1198/jasa.2009.ap08629.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. In: *PLoS genetics* 4.8, e1000167. DOI: 10.1371/journal.pgen.1000167.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf (2012). *Statistical Disclosure Control*. Wiley.

Jansen, B. J., S.-g. Jung, and J. Salminen (2023). Employing Large Language Models in Survey Research. In: *Natural Language Processing Journal* 4, p. 100020. ISSN: 2949-7191. DOI: 10.1016/j.nlp.2023.100020.

Kamat, G. and R. Gutman (2026). Analysis of Linked Files: A Missing Data Perspective. In: *Statistical Science* 41.1, pp. 28–48. DOI: 10.1214/24-STS939.

Kapania, S., S. Ballard, A. Kessler, and J. W. Vaughan (2025). Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 16 pages. DOI: 10.1145/3715275.3732005.

Kinney, S. K., J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. In: *International Statistical Review* 79.3, pp. 362–384. DOI: 10.1111/j.1751-5823.2011.00153.x.

Lin, S., M. Bun, M. Gaboardi, E. D. Kolaczyk, and A. Smith (2024). Differentially private confidence intervals for proportions under stratified random sampling. In: *Electronic Journal of Statistics* 18.1, pp. 1455–1494. DOI: 10.1214/24-EJS2234.

Lin, S., E. Paquette, and E. D. Kolaczyk (July 2024). Differentially Private Linear Regression With Linked Data. In: *Harvard Data Science Review* 6.3.

Lin, S., A. Slavković, and D. R. Bhoomireddy (2026). Differentially Private Linear Regression and Synthetic Data Generation with Statistical Guarantees. In: *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research 300.

Lohr, S. L. (2021). *Sampling: Design and Analysis*. 3rd ed. CRC Press.

Matthews, G. J. and O. Harel (2011). Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy. In: *Statistics Surveys* 5, pp. 1–29.

Mellon, J., J. Bailey, R. Scott, J. Breckwoldt, M. Miori, and P. Schmedeman (Jan. 2024). Do AIs Know What the Most Important Issue Is? Using Language Models to Code Open-Text Social Survey Responses at Scale. In: *Research & Politics* 11.1. DOI: 10.1177/20531680241231468.

Narayanan, A. and V. Shmatikov (2008). "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy*, pp. 111–125. DOI: 10.1109/SP.2008.33.

Reiter, J. P. (2005). Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168.1, pp. 185–205. DOI: 10.1111/j.1467-985X.2004.00343.x.

Rothschild, D. M., T. D. Buskirk, S. Eckman, D. S. Hillygus, F. Kreuter, and D. Lazer (2025). Successfully Navigating the Disruption AI Will Bring to Survey Research. In: *The Survey Statistician* 92, pp. 30–44.

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. In: *Journal of Official Statistics* 9.2, pp. 461–468.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer.

Seeman, J. (2023). “Theoretical and Applied Problems in Partially Private Data”. Doctoral Dissertation. University Park, PA: The Pennsylvania State University. URL: https://etda.libraries.psu.edu/catalog/23008jhs5496.

Seeman, J., M. Reimherr, and A. Slavković (2022). Formal Privacy for Partially Private Data. In: *ArXiv* abs/2204.01102. URL: https://arxiv.org/abs/2204.01102.

Seeman, J., Y. Si, and J. P. Reiter (2026). “Differentially Private Population Quantity Estimates via Survey Weight Regularization”. In: *Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and New Findings*. University of Chicago Press. Chap. 6. URL: https://www.nber.org/books-and-chapters/data-privacy-protection-and-conduct-applied-research-methods-approaches-and-new-findings/differentially-private-population-quantity-estimates-survey-weight-regularization.

Shokri, R., M. Stronati, C. Song, and V. Shmatikov (May 2017). “Membership Inference Attacks Against Machine Learning Models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. DOI: 10.1109/SP.2017.41.

Skinner, C. J. (2009). “Statistical Disclosure Control for Survey Data”. In: *Handbook of Statistics*. Vol. 29. Elsevier, pp. 381–396.

Slavkovic, A. and J. Lee (May 2010). Synthetic two-way contingency tables that preserve conditional frequencies. In: *Statistical Methodology* 7, pp. 225–239. DOI: 10.1016/j.stamet.2009.11.002.

Slavković, A. and J. Seeman (2023). Statistical Data Privacy: A Song of Privacy and Utility. In: *Annual Review of Statistics and Its Application* 10, pp. 189–218. DOI: 10.1146/annurev-statistics-033121-112921. URL: https://doi.org/10.1146/annurev-statistics-033121-112921.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. In: *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05, pp. 557–570.

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. In: *arXiv preprint arXiv:2304.06588*.

Wang, Y.-X. (2018). Revisiting Differentially Private Linear Regression: Optimal and Adaptive Prediction & Estimation in Unbounded Domain. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Webb, K., P. Protivash, J. Durrell, D. Toth, A. Slavković, and D. Kifer (2026). Statistics-Friendly Confidentiality Protection for Establishment Data, with Applications to the QCEW. In: *PoPETS*. To appear. arXiv:2509.01597. arXiv: 2509.01597 [cs.CR]. URL: https://arxiv.org/abs/2509.01597.

Woo, Y. M. J. and A. B. Slavkovic (2012). “Logistic Regression with Variables Subject to Post Randomization Method”. In: *Privacy in Statistical Databases*, pp. 116–130. URL: https://api.semanticscholar.org/CorpusID:5871044.

Wuttke, A., M. Aßenmacher, C. Klamm, M. M. Lang, Q. Würschinger, and F. Kreuter (May 2025). AI Conversational Interviewing: Transforming Surveys with LLMs as Adaptive Interviewers. In: *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* LaTeCH-CLfL 2025, pp. 179–204. DOI: 10.18653/v1/2025.latechclfl-1.17.

Xiao, Z., M. X. Zhou, Q. V. Liao, G. Mark, C. Chi, W. Chen, and H. Yang (June 2020). Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-Ended Questions. In: *ACM Transactions on Computer-Human Interaction* 27.3. ISSN: 1073-0516. DOI: 10.1145/3381804.