*Journal of Survey Statistics and Methodology (2022) **10**, 596–617*

IMPROVING THE UTILITY OF POISSON-DISTRIBUTED, DIFFERENTIALLY PRIVATE SYNTHETIC DATA VIA PRIOR PREDICTIVE TRUNCATION WITH AN APPLICATION TO CDC WONDER

HARRISON QUICK *

CDC WONDER is a web-based tool for the dissemination of epidemiologic data collected by the National Vital Statistics System. While CDC WONDER has built-in privacy protections, they do not satisfy formal privacy protections such as differential privacy and thus are susceptible to targeted attacks. Given the importance of making high-quality public health data publicly available while preserving the privacy of the underlying data subjects, we aim to improve the utility of a recently developed approach for generating Poisson-distributed, differentially private synthetic data by using publicly available information to truncate the range of the synthetic data. Specifically, we utilize county-level population information from the US Census Bureau and national death reports produced by the CDC to inform prior distributions on county-level death rates and infer reasonable ranges for Poisson-distributed, county-level death counts. In doing so, the requirements for satisfying differential privacy for a given privacy budget can be reduced by several orders of magnitude, thereby leading to substantial improvements in utility. To illustrate our proposed approach, we consider a dataset comprised of over 26,000 cancer-related deaths from the Commonwealth of Pennsylvania belonging to over 47,000 combinations of cause-of-death and demographic variables such as age, race, sex, and county-of-residence and demonstrate the proposed framework's ability to preserve features such as geographic, urban/rural, and racial disparities present in the true data.

HARRISON QUICK is an Assistant Professor of Biostatistics in the Department of Epidemiology and Biostatistics, Drexel University, Philadelphia, PA 19104, USA.

Funding for this work was provided by the National Science Foundation, NSF-SES-1943730.

*Address correspondence to Harrison Quick, Department of Epidemiology and Biostatistics, Drexel University, Philadelphia, PA 19104, USA; E-mail: hsq23@drexel.edu

https://doi.org/10.1093/jssam/smac007 Advance access publication 4 April 2022

KEYWORDS: Bayesian methods; Cancer mortality; Confidentiality; Data suppression; Disclosure risk; Spatial data.

Statement of Significance

Access to high-quality public-use data is crucial for many fields of research—especially public health—but when deciding *how* to release data for public use, an agency must weigh the potential risks to data subjects of an unintended disclosure of their personal information. This paper proposes an improvement on a recently developed framework for producing synthetic data with provable privacy protections and illustrates the proposed method using a dataset comprised of the number of cancer-related deaths in Pennsylvania counties stratified by a variety of demographic factors.

1. INTRODUCTION

CDC WONDER—the CDC's Wide-ranging Online Data for Epidemiologic Research system—is a web-based tool for the dissemination of epidemiologic data collected by the National Vital Statistics System (Friede, Reid, and Ory, 1993). Via CDC WONDER, researchers can gain immediate access to vital statistics data, such as the number of births and deaths stratified by geographic region (e.g., state, county), demographic variables (e.g., age, race, sex), specific causes of death [based on International Classification of Disease (ICD) codes], and year from 1968 to 2019, subject to the suppression of small counts (CDC, 2003). Unfortunately, suppression techniques like those implemented on CDC WONDER have been shown to be susceptible to targeted attacks (e.g., Dinur and Nissim, 2003; Holan, Toth, Ferreira, and Karr, 2010; Quick, Holan, and Wikle, 2015), motivating the development of alternative approaches to safely disseminate the nation's vital statistics data for public-use.

To this end, recent work has aimed at replacing the existing CDC WONDER with a "Synthetic CDC WONDER" in which all county-level counts would be replaced by *synthetic* values generated in a Bayesian posterior predictive framework. Quick and Waller (2018) proposed the use of models from the disease mapping literature—specifically, the conditional autoregressive (CAR) model of Besag, York, and Mollié (1991) and the multivariate CAR model of Gelfand and Vounatsou (2003)—to generate synthetic counts of the number of heart disease-related deaths in US counties over a 10-year period across multiple age groups. While this approach may produce synthetic data with high utility by virtue of estimating spatial-, temporal-, and between-age sources of dependence in the true data and preserving those dependencies

in the synthetic data, the approach of Quick and Waller (2018) has not been shown to satisfy formal privacy protections such as the definition of *differential privacy* (Dwork, McSherry, Nissim, and Smith, 2006).

In contrast, Quick (2021) set out to create a differentially private framework for generating synthetic data in the context of CDC WONDER. Inspired by the work of Machanavajjhala, Kifer, Abowd, Gehrke, and Vilhuber (2008)—the methodological framework behind the US Census Bureau's OnTheMap tool—Quick (2021) established criteria in which synthetic data generated from the posterior predictive distribution of a Poisson-gamma model could satisfy ϵ-differential privacy. Specifically, the approach assumes the number of events in a given region arises from a Poisson distribution (as is common in the disease mapping literature; e.g., Brillinger, 1986) and incorporates external information regarding the size of the at-risk population and an estimate of the event rate in an effort to produce synthetic data with greater utility. Furthermore, Quick (2021) demonstrates how the approach of Machanavajjhala et al. (2008) can be viewed as a special case of the proposed Poisson-gamma framework in which the underlying population sizes and event rates are assumed to be equal for all groups.

The drawback of the approach of Quick (2021)—and the multinomial-Dirichlet approach of Machanavajjhala et al. (2008) that inspired it—is that the worst case scenario underlying the criteria for satisfying ϵ-differential privacy is highly unrealistic. In particular, it limits the disclosure risk in the scenario in which a region that experienced just *one* event in the true dataset is assigned *all* of the events in the synthetic data. In the context of data from CDC WONDER, this could imply that the number of deaths allocated to a region could far exceed the size of its at-risk population. In this paper, we propose the use of *prior predictive truncation*—that is, restricting the domain of the synthetic data based on the prior predictive distribution—to improve the utility of our synthetic data. Section 2 begins with a brief overview of how synthetic data can be generated in a posterior predictive fashion and a definition of differential privacy in the context of generating synthetic data. We then proceed to summarize the work of Quick (2021) before shifting the focus to our proposed prior predictive truncation approach. To illustrate the proposed methods, we consider a dataset from 1980 comprised of over 26,000 cancer-related deaths from over 47,000 demographic strata from the Commonwealth of Pennsylvania—these data and the external information we consider "publicly available" are described in section 3.1. Synthetic data are then generated under various modeling assumptions and various levels of privacy protections in section 3.2. Following a brief overview of the paper's key findings, we discuss the work's implications in the context of a "Synthetic CDC WONDER" and next steps in section 4.

2. METHODS

2.1 Notation and Background

We let y_i denote the number of events belonging to group i out of a population of size n_i, for $i = 1, \ldots, I$ and $I \geq 2$. While each individual y_i is deemed potentially sensitive, we assume $y_\cdot = \sum_i y_i > 0$ is not sensitive and thus is publicly available; for example, annual reports released by the CDC include the number of deaths due to major causes of death at the *state* level (e.g., deaths due to cancer), but not at the county level (e.g., table 12 of Kochanek, Murphy, Xu, and Arias, 2019). Furthermore, while the presentation used here indexes the data by a single subscript, in many settings (including that used in section 3), it may be more natural to include *multiple* subscripts to denote multiple subgroups (e.g., age, race, and sex).

Before describing the framework of Quick (2021) and our approach for prior predictive truncation, we begin by describing the general framework we use to draw synthetic data, $\mathbf{z}$, from the posterior predictive distribution that satisfy ϵ-differential privacy (Dwork et al., 2006). First, we specify a distribution for the true data, $\mathbf{y} = (y_1, \ldots, y_I)^T$, given a collection of model parameters, ϕ, denoted $p(\mathbf{y}|\phi)$. We then specify a prior distribution for ϕ given a set of known hyperparameters, $\boldsymbol{\psi}$, denoted $p(\phi|\boldsymbol{\psi})$ and obtain the posterior distribution $p(\phi|\mathbf{y}, \boldsymbol{\psi})$. From this, we can then sample $\mathbf{z}$ from the posterior predictive distribution, $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\psi}) = \int p(\mathbf{z}|\phi)p(\phi|\mathbf{y}, \boldsymbol{\psi})d\phi$. The data synthesis mechanism $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\psi})$ is said to be ϵ-differentially private if for all possible $\mathbf{y}$ and $\mathbf{z}$ and for any hypothetical dataset $\mathbf{x} = (x_1, \ldots, x_I)^T$ with $||\mathbf{x} - \mathbf{y}||_1 = 2$ and $\sum_i x_i = \sum_i y_i$—that is, where there exists i and i' such that $x_i = y_i - 1$ and $x_{i'} = y_{i'} + 1$ with all other values equal—then

$$\left| \log \frac{p(\mathbf{z}|\mathbf{y}, \psi)}{p(\mathbf{z}|\mathbf{x}, \psi)} \right| \leq \epsilon. \tag{1}$$

The objective of the expression in (1) is to formally measure and restrict the amount of information about the true data, $\mathbf{y}$, that is, "leaked" via the release of the synthetic data, $\mathbf{z}$. The parameter ϵ is thus often referred to as the "privacy budget," with a small privacy budget (e.g., $\epsilon < 1$) implying that you cannot "afford" to leak information about the data.

The US Census Bureau's OnTheMap tool was the first synthetic data production system based on differential privacy. OnTheMap is a tool that provides access to synthetic data on commuting patterns of US workers—for example, the number of residents of geographic region A who commute to work in geographic region B, denoted $y_{A;B}$. These synthetic data are generated using a multinomial-Dirichlet mechanism proposed by Machanavajjhala et al. (2008) which, under certain conditions, can satisfy (1). Specifically, the approach assumes that the number of workers commuting to region B from each of the various regions in the spatial domain, denoted $\mathbf{y}_B = (y_{1;B}, \ldots, y_{I;B})^T$, follows a

multinomial distribution with $\sum_i y_{i;B} = y_{B\cdot}$ total commuters and probabilities denoted by $\theta_{i;B}$, and where $\boldsymbol{\theta}_B \sim \text{Dir}(\boldsymbol{\alpha}_B)$; synthetic data, $\mathbf{z}_B$, are then sampled from the resulting posterior predictive distribution. For $p(\mathbf{z}_B|\mathbf{y}_B, \boldsymbol{\alpha}_B)$ to satisfy differential privacy for a given privacy budget, $\epsilon > 0$, the hyperparameters $\alpha_{i;B}$ must be sufficiently large. When ϵ is small, however, the requirements for the $\alpha_{i;B}$ become prohibitively high, resulting in a prior distribution which would dominate the data and thereby hinder the utility of the synthetic data (Charest, 2011). The approach of Machanavajjhala et al. (2008) can be viewed as a special case of the approach of Quick (2021), and thus a more detailed derivation of its properties will be discussed in the following subsection.

While the Census Bureau's first foray into differential privacy took the form of a posterior predictive sampling approach, the "TopDown" algorithm developed by the Census Bureau for the 2020 Decennial Census of Population and Housing (Abowd, Ashmead, Simson, Kifer, Leclerc, et al., 2019) takes a more conventional approach in which differentially private noise is added to the true values—sometimes referred to as *output perturbation* (Dwork et al., 2006; Ghosh, Roughgarden, and Sundararajan, 2012). As the TopDown algorithm and other output perturbation-based approaches are fundamentally different than the work proposed in Machanavajjhala et al. (2008) and Quick (2021), this paper's focus is to improve the utility of the Poisson-gamma framework, leaving more thorough comparisons of disparate differentially private techniques for future research.

2.2 The Poisson-Gamma Mechanism

When generating synthetic data in the context of CDC WONDER—for example, the number of deaths due to a given cause of death in a given county—there are two important factors that ought to be taken into account. First and foremost, US counties vary wildly in their population sizes, even within the same state—for example, the most populated county in the state of Texas (Harris County) is home to more than 4.7 million residents while the three least populated counties have fewer than 500 residents each. In addition, death rates can vary substantially by cause of death and by demographic factors like age, race/ethnicity, and sex. Thus, a synthetic data mechanism with the flexibility to account for heterogeneity in population sizes and/or event rates should be expected to outperform an otherwise comparable mechanism that fails to do so. From this point forward, we assume information regarding group-specific population sizes, n_i, and suitable estimates of the underlying event rates are publicly available—support for this assumption will be provided as part of the illustrative example in section 3. Implications for situations where estimates of the population sizes and/or event rates are themselves subject to privacy protections are discussed in section 4.

With this setup in mind, we follow the approach of Quick (2021) and let

$$y_i|\lambda_i \sim \text{Pois}(n_i\lambda_i) \text{ and } \lambda_i \sim \text{Gamma}(a_i, b_i), \tag{2}$$

where λ_i denotes the underlying event rate in group i and a_i and b_i denote group-specific hyperparameters such that $E[\lambda_i|a_i, b_i] = a_i/b_i = \lambda_{i0}$ is known (i.e., based on publicly available information). It is then straightforward to show that $\lambda_i|y_i \sim \text{Gamma}(y_i + a_i, n_i + b_i)$. In addition, since the λ_i are conditionally independent, any weighted average of the λ_i with unnormalized weights $n_i + b_i$ will also be a gamma random variable—for example, the weighted average of event rates *not* associated with group i, denoted $\lambda_{(i)}$, can be written as

$$\lambda_{(i)}|\mathbf{y}_{(i)} \sim \text{Gamma}(y_{(i)} + a_{(i)}, n_{(i)} + b_{(i)}), \text{ where } \lambda_{(i)} = \frac{\sum_{j\neq i}(n_j + b_j)\lambda_j}{\sum_{j\neq i}(n_j + b_j)} \tag{3}$$

and $y_{(i)} = \sum_{j\neq i} y_j$, with similar definitions for $a_{(i)}$, $n_{(i)}$, and $b_{(i)}$. From (2) and (3), it can then be shown that the respective posterior predictive distributions for the synthetic counts, z_i and $z_{(i)}$, will each take the form of a negative binomial distribution—for example, $z_i|y_i, a_i, b_i \sim \text{NegBin}(y_i + a_i, n_i/(b_i + 2 \times n_i))$. Then, if we wish to generate a vector of synthetic counts, $\mathbf{z}_i = (z_i, z_{(i)})^T$ such that $z_i + z_{(i)} = y_.$, our interest lies in the joint distribution, $p(\mathbf{z}_i|\mathbf{y}, \mathbf{a}, \mathbf{b})$, which can be expressed as

$$p(\mathbf{z}_i|\mathbf{y}, \mathbf{a}, \mathbf{b}) = \frac{\frac{\Gamma(z_i+y_i+a_i)}{z_i!}\left(\frac{n_i}{b_i+2\times n_i}\right)^{z_i} \times \frac{\Gamma(z_{(i)}+y_{(i)}+a_{(i)})}{z_{(i)}!}\left(\frac{n_{(i)}}{b_{(i)}+2\times n_{(i)}}\right)^{z_{(i)}}}{\sum_{z=0}^{y_.}\frac{\Gamma(z+y_i+a_i)}{z!}\left(\frac{n_i}{b_i+2\times n_i}\right)^{z} \times \frac{\Gamma(y_.-z+y_{(i)}+a_{(i)})}{(y_.-z)!}\left(\frac{n_{(i)}}{b_{(i)}+2\times n_{(i)}}\right)^{(y_.-z)}}. \tag{4}$$

Thus, demonstrating that $p(\mathbf{z}_i|\mathbf{y}, \mathbf{a}, \mathbf{b})$ satisfies ϵ-differential privacy requires that we can establish bounds for the ratio

$$\frac{p(\mathbf{z}_i|\mathbf{y}, \mathbf{a}, \mathbf{b})}{p(\mathbf{z}_i|\mathbf{x}, \mathbf{a}, \mathbf{b})} = \frac{C(\mathbf{x}, \mathbf{n}, \mathbf{a}, \mathbf{b})}{C(\mathbf{y}, \mathbf{n}, \mathbf{a}, \mathbf{b})} \times \frac{\Gamma(z_i + y_i + a_i)}{\Gamma(z_i + x_i + a_i)} \times \frac{\Gamma(z_{(i)} + y_{(i)} + a_{(i)})}{\Gamma(z_{(i)} + x_{(i)} + a_{(i)})}, \tag{5}$$

where $\mathbf{x}_i = (x_i, x_{(i)})^T$ represents a hypothetical dataset such that $\|\mathbf{x}_i - \mathbf{y}_i\|_1 = 2$ and where

$$C(\mathbf{y}, \mathbf{n}, \mathbf{a}, \mathbf{b}) = \sum_{z=0}^{y_.} \frac{\Gamma(z + y_i + a_i)}{z!} \frac{\Gamma(y_. - z + y_{(i)} + a_{(i)})}{(y_. - z)!} \times r_i(\mathbf{n}, \mathbf{b})^z,$$

and $r_i(\mathbf{n}, \mathbf{b}) = (b_{(i)}/n_{(i)} + 2)/(b_i/n_i + 2)$. As proven in Quick (2021)—and as outlined in appendix A.1 of the supplementary materials file online—in order for (4) to satisfy ϵ-differential privacy, we require

$$a_i \geq \frac{y_\cdot}{e^\epsilon / \nu_i - 1}, \text{ where } \nu_i = \frac{y_\cdot \times \lceil 1 - r_i(\mathbf{n}, \mathbf{b}) \rceil^+ + a_{(i)} + y_\cdot - 1}{a_{(i)} + y_\cdot - 1}. \quad (6)$$

Furthermore, if (6) holds for all i simultaneously, then the mechanism imposed by (2) for $\mathbf{z} = (z_1, \ldots, z_I)$ for $I > 1$ under the constraint that $\sum_i z_i = y_\cdot$, denoted $p(\mathbf{z}|\mathbf{y}, \mathbf{a}, \mathbf{b})$, will satisfy ϵ-differential privacy. To sample from $p(\mathbf{z}|\mathbf{y}, \mathbf{a}, \mathbf{b})$, we can leverage the fact that because the y_i are conditionally independent given $\boldsymbol{\lambda}$, we can write

$$\mathbf{y}|y_\cdot, \boldsymbol{\lambda} \sim \text{Mult}(y_\cdot, \boldsymbol{\pi}), \text{ where } \pi_i = \frac{n_i \lambda_i}{\sum_{j=1}^{I} n_j \lambda_j}. \quad (7)$$

Then, if λ_i^* is drawn from the posterior distribution $\lambda_i|y_i \sim \text{Gamma}(y_i + a_i, n_i + b_i)$ and if π_i^* is defined as in (7) as a function of λ_i^*, then sampling $\mathbf{z}|\boldsymbol{\pi}^*, y_\cdot \sim \text{Mult}(y_\cdot, \boldsymbol{\pi}^*)$ will be equivalent to a draw from $p(\mathbf{z}|\mathbf{y}, \mathbf{a}, \mathbf{b})$.

If $n_i = n$ and $E[\lambda_i|a_i, b_i] = \lambda_0$ for all i—that is, if we assume all groups have the same population size and the same prior expected event rate—then the Poisson-gamma mechanism of Quick (2021) is *mathematically equivalent* to the multinomial-Dirichlet mechanism of Machanavajjhala et al. (2008) with $\boldsymbol{\alpha} = \mathbf{a}$. While the Poisson-gamma mechanism addresses two key limitations of the multinomial-Dirichlet framework of Machanavajjhala et al. (2008)—namely that it allows for heterogeneity in population sizes via the n_i parameters and heterogeneity in the prior event rates via the hyperparameters, a_i and b_i—one drawback both approaches share is that they are both functions of the total number of events, $y_\cdot$, and thus when $y_\cdot$ is large, the restriction in (6) quickly becomes overwhelming. In contrast, the methods proposed in section 2.3 aim to construct a framework in which the informativeness of the prior is primarily a function of $E[y_i|\mathbf{a}, \mathbf{b}]$. Because the hyperparameters and prior predictive expected values, $E[y_i|\boldsymbol{\alpha}] = y_\cdot/I$, under the multinomial-Dirichlet framework are not intended to reflect any prior beliefs regarding the nature of the true data, our methodological focus from this point forward will be restricted to the more general Poisson-gamma framework.

2.3 Prior Predictive Truncation

The reason the Poisson-gamma framework proposed by Quick (2021) suffers from low utility for moderate values of ϵ is that it is designed to protect against an extremely improbable worst case scenario—that is, where *all* of the events in the synthetic dataset are assigned to a group in which only *one* event occurred. To address this issue in the multinomial-Dirichlet setting, Machanavajjhala et al. (2008) proposed a framework referred to as (ϵ, δ)-probabilistic differential privacy in which the synthetic data would satisfy ϵ-

differential privacy with probability $1 - \delta$ for $\delta > 0$. Aside from not satisfying *pure* ($\delta = 0$) differential privacy, a drawback of this approach is that the hyperparameters, $\boldsymbol{\alpha}$, cannot be reported because they are based on the *posterior* predictive distribution and thus are informed by the data.

Here, we propose the use of *prior* predictive truncation, specifically the use of n_i and $\lambda_{i0} = E[\lambda_i | a_i, b_i]$ to specify *a priori* upper and lower bounds on z_i which take the form:

$$L_i = F^{-1}(\alpha/2 | n_i \lambda_{i0}) \leq z_i \leq F^{-1}(1 - \alpha/2 | n_i \lambda_{i0}) = U_i, \tag{8}$$

where $F^{-1}(\cdot|\mu)$ denotes the inverse cdf of a Poisson random variable with mean μ and where $\alpha \in (0, 1/2)$ corresponds to the tail probabilities used for the truncation, with smaller values of α leading to wider bounds; as a rule-of-thumb, we suggest letting $\alpha = \min\{0.001, I^{-1}\}$, where choosing $\alpha = I^{-1}$ when I is large is akin to a Bonferroni correction when conducting multiple comparisons to achieve a low family-wise error rate. It should be emphasized here that the objective of (8) is *not* to force each $z_i \approx y_i$, but simply to restrict z_i to a range of plausible values. Moreover, it should also be emphasized that while it would be *desirable* for $y_i \in [L_i, U_i]$ for all i, this is not a *requirement* as that alone may leak information about $\mathbf{y}$ and thus be a violation of differential privacy. That said, we *must* truncate the posterior contribution of y_i to the range $[L_i, U_i]$—for example, any $y_i > U_i$ will be replaced by U_i in (2). As a result, this approach is highly sensitive to the quality of the prior information—for example, if $y_i \gg E[y_i | \mathbf{a}, \mathbf{b}] = n_i \lambda_{i0}$, then the proposed bounds may be too restrictive and thus may reduce the utility of the synthetic data.

To implement the proposed prior predictive truncation, we begin by assuming that $E[y_i | \mathbf{a}, \mathbf{b}] < E[y_{(i)} | \mathbf{a}, \mathbf{b}]$ for all i—that is, that no single group "dominates" the remaining groups. While this assumption can certainly be violated in practice—for example, when a majority of a state's population belongs to a single urban center—this assumption will tend to hold in scenarios in which the data are stratified by multiple factors (e.g., by spatial region *and* demographic categories like age, race, and sex). Nevertheless, appendix A.2 of the supplementary materials file online outlines how one can proceed when this assumption is violated. Next, in order to demonstrate that this approach can be proven to satisfy ϵ-differential privacy, we first restrict our focus to the "group i versus *not* group i" scenario, which—as demonstrated in (A.8) of appendix A.2 of the supplementary materials file online—yields a ratio of bivariate posterior predictive distributions for $\mathbf{z}_i = (z_i, z_{(i)})^T$ of the form:

$$\frac{p(\mathbf{z}_i | \mathbf{y}, \mathbf{a}, \mathbf{b}, \mathcal{B})}{p(\mathbf{z}_i | \mathbf{x}, \mathbf{a}, \mathbf{b}, \mathcal{B})} = \frac{C_{L_i}^{U_i}(\mathbf{x}, \mathbf{n}, \mathbf{a}, \mathbf{b})}{C_{L_i}^{U_i}(\mathbf{y}, \mathbf{n}, \mathbf{a}, \mathbf{b})} \times \frac{\Gamma(z_i + y_i + a_i)}{\Gamma(z_i + x_i + a_i)} \times \frac{\Gamma(z_{(i)} + y_{(i)} + a_{(i)})}{\Gamma(z_{(i)} + x_{(i)} + a_{(i)})}, \tag{9}$$

where $\mathcal{B} = \{(L_i, U_i) : i = 1, \ldots, I\}$ denotes the set of bounds based on (8) and

$$C_{L_i}^{U_i}(\mathbf{y},\mathbf{n},\mathbf{a},\mathbf{b}) = \sum_{z=L_i}^{U_i} \frac{\Gamma(z+y_i+a_i)}{z!} \times \frac{\Gamma(y_\cdot - z + y_{(i)} + a_{(i)})}{(y_\cdot - z)!} \times r_i(\mathbf{n},\mathbf{b})^z. \tag{10}$$

The key result of this paper is then summarized in theorem 1 below:

Theorem 1. *Let* $\boldsymbol{C}_{\boldsymbol{L_i}}^{\boldsymbol{U_i}}(\mathbf{y},\mathbf{n},\mathbf{a},\mathbf{b})$ *be as defined in (10) and* $\boldsymbol{E}[\boldsymbol{y_i}|\boldsymbol{a_i},\boldsymbol{b_i}] \leq \boldsymbol{E}[\boldsymbol{y_{(i)}}|\boldsymbol{a_{(i)}},\boldsymbol{b_{(i)}}]$. *Then, if* $\mathbf{x} = (\boldsymbol{x_i},\boldsymbol{x_{(i)}})^T$ *and* $\mathbf{y} = (\boldsymbol{y_i},\boldsymbol{y_{(i)}})^T$ *denote vectors of nonnegative integers of length 2 such that* $\boldsymbol{x_i} + \boldsymbol{x_{(i)}} = \boldsymbol{y_i} + \boldsymbol{y_{(i)}} = \boldsymbol{y_\cdot}$, *then*

$$\frac{C_{L_i}^{U_i}(\mathbf{x},\mathbf{n},\mathbf{a},\mathbf{b})}{C_{L_i}^{U_i}(\mathbf{y},\mathbf{n},\mathbf{a},\mathbf{b})} < \frac{y_\cdot - L_i + a_{(i)} + y_{(i)}}{L_i + a_i + y_i - 1}. \tag{11}$$

The upper bound obtained from theorem 1 occurs when $x_i = y_i - 1$ and $x_{(i)} = y_{(i)} + 1$; thus, the expression in (9) can be bounded above by letting $y_i = L_i + 1$ and $z_i = U_i$ to yield

$$\frac{p(\mathbf{z}_i|\mathbf{y},\mathbf{a},\mathbf{b},\mathcal{B})}{p(\mathbf{z}_i|\mathbf{x},\mathbf{a},\mathbf{b},\mathcal{B})} \leq \frac{y_\cdot - L_i + a_{(i)} + y_\cdot - L_i - 1}{L_i + a_i + L_i} \times \frac{U_i + a_i + L_i}{y_\cdot - U_i + a_{(i)} + y_\cdot - L_i - 1}. \tag{12}$$

Thus, for $p(\mathbf{z}_i|\mathbf{y},\mathbf{a},\mathbf{b},\mathcal{B})$ to satisfy ϵ-differential privacy, we require

$$a_i \geq \frac{U_i - L_i}{e^\epsilon/\nu_i - 1} - 2 * L_i, \text{ where } \nu_i = \frac{y_\cdot - L_i + a_{(i)} + y_\cdot - L_i - 1}{y_\cdot - U_i + a_{(i)} + y_\cdot - L_i - 1}. \tag{13}$$

It should be noted that the criteria in (13) is similar—but not *equivalent*—to that from the untruncated framework in (6) when $L_i = 0$ and $U_i = y_\cdot$. This is because unlike the approximation-based bound from (6), the bound from theorem 1 is obtained by replacing the ratio of summations in (11) with a ratio of just the *first* terms of the summation—that is, the $z = L_i$ terms in (10)—which in turn yields a more *conservative* upper bound. See appendix A.2 of the supplementary materials file online for a full derivation of (13) and a proof of theorem 1.

While the requirement in (13) is based on a bivariate set of synthetic data, $\mathbf{z}_i$, requiring that a_i satisfy the appropriate criteria from (13) for all i simultaneously will ensure that the mechanism for generating the complete vector of synthetic data, denoted $p(\mathbf{z}|\mathbf{y},\mathbf{a},\mathbf{b},\mathcal{B})$, will satisfy ϵ-differential privacy. The benefit of this approach is that the various a_i are allowed to vary as a function of $E[y_i|a_i,b_i]$, and thus groups with smaller expected event counts should be expected to receive *far* less informative priors under (13) than under (6), especially as the total number of events, $y_\cdot$, increases. It should also be noted that

because the expression for a_i in (13) is a function of $a_{(i)}$, we require the use of an iterative algorithm to simultaneously calculate the a_i and b_i parameters subject to the model's constraints—see appendix B.1 of the supplementary materials file online for more details. In addition, because the bound from theorem 1 can be quite conservative, an algorithm for optimizing the model's hyperparameters is proposed in appendix B.2 of the supplementary materials file online. Finally, when ϵ is large, the lower bounds on the a_i can be quite small. While this should seemingly correspond to increased utility, work by Kerman (2011) has demonstrated the potential for the posterior distribution for λ_i resulting from (2) to concentrate near zero when $y_i + a_i \approx 0$, which can then lead to a preponderance of zeros in the synthetic data. To overcome this, we follow the guidance of Kerman (2011) and require $a_i > 1/3$ when $L_i = 0$—that is, when our prior information deems that $y_i = 0$ is plausible.

As a simple illustration of how this approach works, we consider a dataset of $y_\cdot = 100$ events belonging to $I = 2$ groups with $E[\mathbf{y}|\mathbf{a}, \mathbf{b}] = (15, 85)^T$ with $\epsilon = 1$. As illustrated by the solid lines in figure 1a, the approach of Quick (2021) maximizes the ratio from (5) when $\mathbf{y} = (1, 99)^T$ and $\mathbf{z} = (100, 0)^T$. While this maximum risk *is* indeed less than the desired threshold of $\exp(\epsilon) = 2.71$, the probability of allocating all $y_\cdot = 100$ events to the group with $E[y_1|\mathbf{a}, \mathbf{b}] = 15$ is extremely low (i.e., less than 4.1×10^{-83}), and protecting against this worst case scenario requires $a_1 > 116$, a value which certainly will overwhelm whatever the true y_1 is. In contrast, the proposed prior predictive truncation approach with $\alpha = 10^{-4}$ maximizes the ratio from (12) for group 1 when $z_1 = 30$, as demonstrated by the dashed lines in figure 1a. It should be noted here that $z_1 = 30$ is twice as large as $E[y_1|a_1, b_1]$ and thus is still a conservative bound—that is, a bound that is not intended to meaningfully restrict the range of z_1. Furthermore, the benefit of this approach is that we now only require $a_1 > 6.85$ to satisfy ϵ-differential privacy. Similarly, the requirement for a_2 is reduced from $a_2 > 58$ to just $a_2 > 0.001$. As a result, we obtain posterior distributions for each λ_i that are far more reliant on the true data than under Quick (2021), thereby producing synthetic data that have greater utility without sacrificing data privacy, as shown in figure 1b. While the results shown here are based on the *optimized* approach for identifying bounds on $\mathbf{a}$, a comparison between the optimized bounds and those based solely on (13) is shown in appendix B.2.4 of the supplementary materials file online.

3. ILLUSTRATIVE ANALYSIS

3.1 Pennsylvania Cancer Death Data

To illustrate the benefits of the prior predictive truncation described in section 2 in a real-world application, we consider a dataset comprised of cancer-related

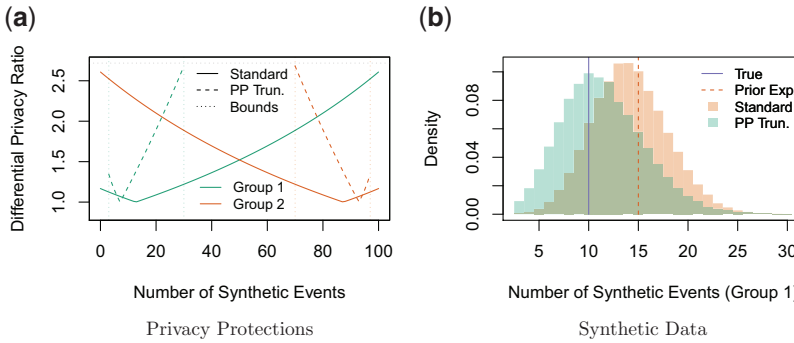

Figure 1. Generating Synthetic Data with $\epsilon = 1$ for $y_{\cdot} = 100$ Events and $\boldsymbol{E}[\mathbf{y}|\mathbf{a},\mathbf{b}] = (\mathbf{15},\mathbf{85})^{T}$. In (a), the prior predictive truncation approach (dashed lines) yields less informative priors, causing the differential privacy ratios from (13) to increase at a much faster rate toward their bounds from (9) (dotted lines) than under the standard framework (solid lines) based on (5) from Quick (2021). Panel (b) compares the posterior predictive distribution for group 1 from the competing approaches for sampling synthetic data—for reference, the true value of $\boldsymbol{y}_1 = 10$ and the prior expected value of $\boldsymbol{E}[\boldsymbol{y}_1|\mathbf{a},\mathbf{b}] = 15$ are highlighted. (a) Privacy protections. (b) Synthetic data.

death counts and population estimates from the Commonwealth of Pennsylvania (PA) in 1980. We let y_{icars} denote the number of deaths in county i (of 67 counties) due to cancer type c (of 9 types) from individuals belonging to age group a (of 13 age groups), race r (white, black, other), of sex s (male/female) (see table 1 for more information). In total, there were $y_{\cdot} = \sum_{icars} y_{icars} = 26,116$ cancer-related deaths in PA in 1980 belonging to these $67 \times 13 \times 9 \times 3 \times 2 = 47,034$ strata, and because these data are from prior to 1989, they are publicly available via CDC WONDER free of suppression; see appendix D of the supplementary materials file online for instructions for accessing these data. Here, we assume that the total number of cancer-related deaths in PA, $y_{\cdot}$, is to be kept invariant (i.e., is safe to disseminate) but that the county-level death counts, y_{icars}, are to be considered sensitive.

The external information used in this analysis comes from two sources. First and foremost, we incorporate the 1980 *bridged-race population estimates*—denoted n_{iars}—released annually by the National Center for Health Statistics (NCHS). As described by Ingram, Parker, Schenker, Weed, Hamilton, et al. (2003), these estimates are model based and are the product of a collaborative agreement between the CDC and the Census Bureau that relies on census data and information from the CDC's National Health Interview Survey. These estimates were developed for the express purpose of creating denominators for vital rates by external researchers, thus we believe it is reasonable to assume that similar estimates will continue to be produced and publicly disseminated despite the Census Bureau's shift to using differential privacy to protect the full

Table 1. Overview of the Structure of the Pennsylvania Cancer Data

Attribute	Levels
County	$i = 1, \ldots, 67$ counties in Pennsylvania
Cancer type	$c = 1, \ldots, 9$ forms of cancer; cancers of the lip, oral cavity, and pharynx (ICD-9: 140–149); cancers of the digestive organs and peritoneum (ICD-9: 150–159); cancers of the respiratory and intrathoracic organs (ICD-9: 160–165); cancers of the breast (ICD-9: 174–175); cancers of the genital organs (ICD-9: 179–187); cancers of the urinary organs (ICD-9: 188–189); cancers of all other and unspecified sites (ICD-9: 170–173, 190–199); leukemia (ICD-9: 204–208); and all other cancers of the lymphatic and hematopoietic tissues (ICD-9: 200–203)
Age	$a = 1, \ldots, 13$ levels; ages <1; ages 1–4; ages 5–9; ages 10–14; ages 15–19; ages 20–24; ages 25–34; ages 35–44; ages 45–54; ages 55–64; ages 65–74; ages 75–84; and ages 85 and older
Race	$r = 1, \ldots, 3$ levels (black, white, and others)
Sex	$s = 1, 2$ levels (male and female)

NOTE.—Cancer types are identified by their International Classification of Diseases, Ninth Revision (ICD-9) codes.

2020 Decennial Census of Population and Housing (Abowd, 2018). That said, it should be acknowledged that producing the bridged-race population estimates using differentially private census data may result in a reduction in their accuracy.

The second source of data used in this analysis comes from the CDC's 1980 National Vital Statistics Report (NCHS, 1985). Specifically, in addition to assuming that $y_{\cdot}$ is known, we assume that the cancer-related death rates *at the national level* for each of the aforementioned strata—that is, the national death rate for each combination of age, race, sex, and form of cancer, denoted $\lambda_{cars;0}$—are publicly known. To put these assumptions into perspective with regards to more recent reports, analogous state-level totals for major causes of death (e.g., cancer) are provided in table 12 of the CDC's 2018 report (Murphy, Xu, Kochanek, Arias, and Tejada-Vera, 2021), along with national death rates by age and selected causes in table 7 of the report and national death rates by race, sex, and selected causes in table 8 of the report. Previous iterations of the report (e.g., those prior to the inception of CDC WONDER) included death rates by age, race, sex, and selected causes (e.g., tables 1–8 and 1–9 of NCHS, 1985)—here, we operate under the assumption that the decision to discontinue publishing tables by age, race, sex, and selected causes was due to the inception of CDC WONDER (where such granular data is available) rather than due to privacy concerns. As such, our goal here is to illustrate how data sources that were—and still are—publicly available can be used to help

produce high-quality synthetic data while still offering formal privacy protections for the more sensitive underlying data. Moreover, should the CDC determine that these national death reports must also be differentially private, we expect that the large denominators underlying these rates would provide stability to the sanitized values. Nevertheless, we discuss this scenario at greater length in section 4.

Figure 2 demonstrates the utility of the prior information used in this analysis. In figure 2a, we compare the age/cause-specific death rates at the national level from 1980 and those at the state level, and in figure 2b, we compare the true death counts to the expected death counts based on the bridged-race population estimates and the national death rates. In both cases, we see a high degree of agreement, as is desired. That said, the decision to *use* this information must be made *prior* to utility checks of this form, as to do otherwise would result in a violation of differential privacy. A discussion of how to choose sources of prior information and the implications on privacy protections is provided in section 4.

3.2 Generating Synthetic Death Counts

To generate differentially private synthetic data, we consider two approaches: the untruncated Poisson-gamma framework of Quick (2021) and the truncated Poisson-gamma framework proposed in section 2.3 with $\alpha = 1/47,034$. For each approach, $L = 1,000$ sets of synthetic data were generated for $\epsilon \in [0.01, 4]$, allowing us to assess the sampling distribution of the synthetic data across a broad range of privacy budgets. Synthetic data generated from the untruncated approach of Quick (2021) were sampled using the approach outlined in section 2.2, and synthetic data generated from the proposed prior predictive truncation approach were sampled using the algorithm outlined in appendix C of the supplementary materials file online. All point and interval estimates presented here are (unless otherwise noted) based on quantiles from the sampling distribution based on the L sets of synthetic data for each level of ϵ—that is, point estimates are based on medians while interval estimates are based on the 2.5 and 97.5 percentiles of the sampling distribution (henceforth referred to as the 95 percent sampling interval).

Before diving into a comparison of the synthetic data themselves, we first compare the prior distributions underlying each of the synthesizers. When using the approach proposed by Quick (2021), the prior distribution for all of our y_{icars} required an informativeness of $a_{icars} > y_{\cdot}/\left(\exp(\epsilon) - 1\right)$. For instance, with $y_{\cdot} = 26,116$ total deaths and $\epsilon = 1$, this implies $a_{icars} > 15,000$. To put this in perspective, the largest value in our dataset is just $\max(y_{icars}) = 237$. In contrast, using the prior predictive truncation approach described in section 2.3, our priors all have $a_{icars} < 17$ with a median value of 0.58, values comparable to the untruncated approach with $\epsilon > 7$. As a result, we should expect the

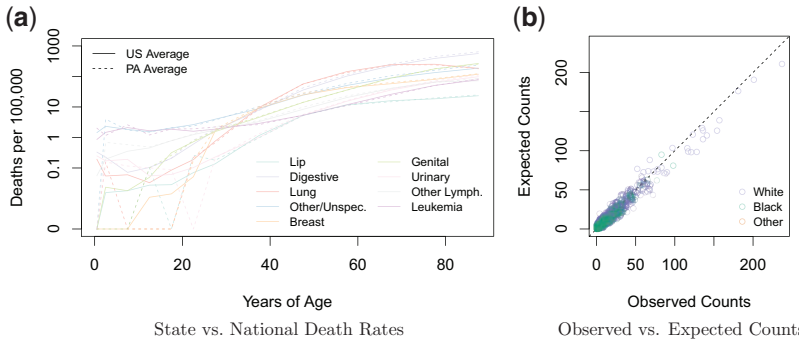

Figure 2. Cause-Specific Death Rates at the National Level and for the State of Pennsylvania. National-level rates are used as prior information for estimating the proper allocation of deaths at the state and county level. (a) State versus national death rates. (b) Observed versus expected counts.

prior predictive truncation approach to put more weight on the observed data, thereby producing synthetic data with greater utility. In addition, it should be noted that lower bounds for the a_{icars} when $\epsilon > 2$ become quite small (e.g., see figures E.1–E.2 of the supplementary materials file online) thus further increases in the privacy budget will have diminishing returns with respect to utility. Finally, we also note that (while not required) none of the y_{icars} fell outside the bounds of the prior predictive truncation used in this analysis, suggesting that the prior information used is highly relevant and that the bounds constructed from our choice of α were sufficiently conservative.

As an initial evaluation of the utility of the synthetic data, figure 3 compares the age-adjusted death rates due to cancer (averaged over all L sets of synthetic data) of both approaches for $\epsilon = 1$ and $\epsilon = 4$. Immediately, we see the effect of the highly informative priors imposed by the untruncated approach of Quick (2021), as the rates in all 67 counties have been pulled toward a common value of 217 deaths per 100,000—that is, the national age-adjusted death rate—and thus that the geographic variation in rates that are present in the true data has all but been smoothed away. In contrast, the synthetic data based on the prior predictive truncation approach largely preserve geographic disparities in rates when $\epsilon = 4$ and maintain *some* heterogeneity in the rates when $\epsilon = 1$. This is more apparent in figure 4, which compares the map of the true age-adjusted rates to the averages based on the synthetic data under the proposed prior predictive truncation approach when $\epsilon = 1$. Here, we see that while the synthetic data have more moderate rates throughout the less populated, rural counties, rates in PA's population centers are allowed to deviate from the norm. Maps for other values of ϵ for the truncated model are provided in figure E.3 of the supplementary materials file online. Due to the disparate levels of informativeness between the truncated and untruncated models (which can best be

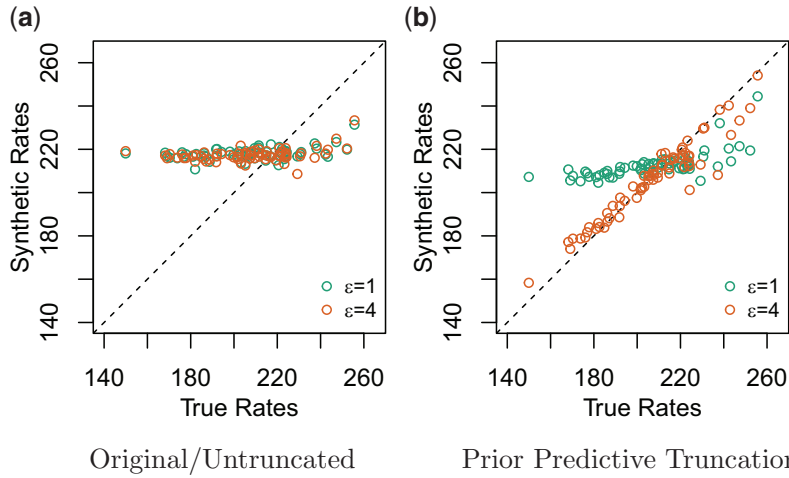

Figure 3. Comparison of Age-Adjusted Cancer-Related Death Rates Based on the Two Approaches for Generating Synthetic Data for $\epsilon = 1$ and $\epsilon = 4$.

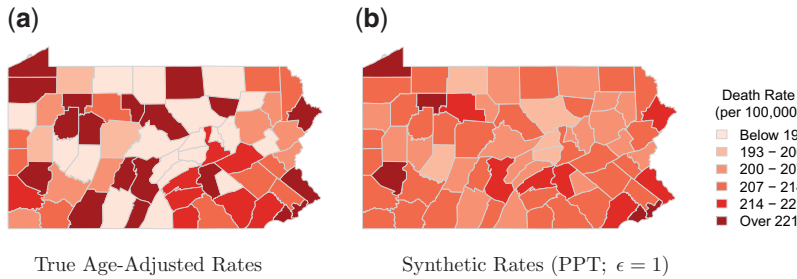

Figure 4. Maps of the Age-Adjusted Cancer Death Rates Based on the True Data and Synthetic Data Generated under the Proposed Prior Predictive Truncation Approach with $\epsilon = 1$.

observed in figure E.1 of the supplementary materials file online), from this point forward we restrict our focus to the results from the proposed prior predictive truncation approach.

We now shift our focus to two commonly investigated disparities in rates: disparities in rates between urban and rural counties and racial disparities. For our assessment of urban/rural disparities, we begin by defining a county as being "urban" if its population density exceeded 245 persons per square mile—that is, having a population density greater than the statewide average—and "rural" otherwise. Based on this definition, 21 of PA's 67 counties are deemed

"urban," and the age-adjusted cancer death rates in these counties were 12 percent higher than those in rural counties. In contrast, because the CDC's annual death reports *do not* break down death rates by urban/rural status, our prior information assumes *no disparities* exist between urban and rural counties. The result of these two competing phenomena can be seen in the urban/rural disparity estimates produced by the synthetic data shown in figure 5a. In particular, while the synthetic data yield accurate estimates of the urban/rural disparity when ϵ is large, the estimates gradually shift toward the null value from the prior as ϵ decreases and more informative priors are used. In addition, note that more informative priors correspond to priors (and posterior predictive distributions) with lower variability; thus as ϵ decreases, the variability in the sampling distribution decreases as well. This is in contrast to more conventional approaches for satisfying differential privacy which increase the level of variability to provide stronger privacy protections; this phenomenon is discussed in more detail in section 4 of Quick (2021).

Similar results are obtained in an investigation of the black/white disparities in cancer death rates, as shown in figure 5b. In the true death data, cancer death rates for black men were nearly 45 percent higher than for their white counterparts, while rates for black women were 15 percent higher than for white women. Unlike the urban/rural disparities example, however, our prior information based on the CDC's annual death reports *do* contain information on racial disparities at the national level. By coincidence, the black/white disparity in cancer death rates among women in Pennsylvania in 1980 was nearly identical to that at the national level, thus in this case our prior information effectively equals the true value. It should be noted that this is *not* to be expected, nor does it imply that the racial disparities at the *county level* are also equivalent to those at the national level. Thus, the takeaway message from figure 5 should not be that the prior predictive truncation approach will *preserve* inference on quantities such as urban/rural disparities or racial disparities, but rather that the synthetic data should be expected to produce estimates between the value from the true data (which will be unknown to data-users) and the publicly available national estimates.

Figure 6 offers two more nuanced illustrations of the Poisson-gamma mechanism's behavior and the importance of disclosing the prior information used in the production of the synthetic data. In figure 6a, we explore the estimated black/white disparity in the rate of death from digestive cancer (ICD-9: 150–159) among women. Here, the estimated black/white disparity *increases* as ϵ decreases because the disparity at the national level is greater than that in the state. Thus, in this scenario, a careful end-user will want to compare the estimate they obtained against the known national average and the null of no disparity prior to drawing any conclusions. This phenomenon is even more clear in figure 6b, which displays the black/white disparity in rates of death from other lymphatic cancer (ICD-9: 200–203) among men. Here, the disparity in the Commonwealth of Pennsylvania indicates higher rates among white men, a

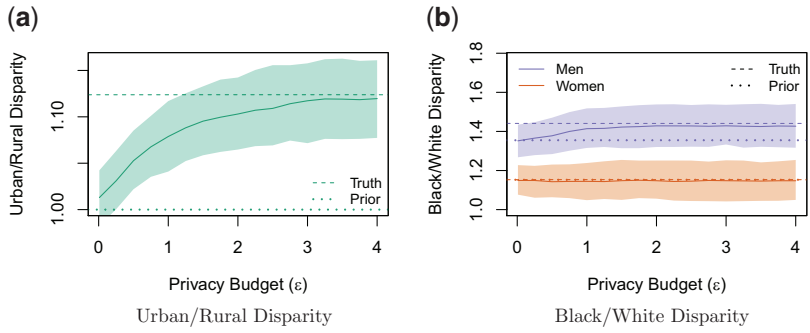

Figure 5. Estimated Urban/Rural Disparities and Black/White Disparities (by Sex) Based on the Synthetic Data Generated from the Posterior Predictive Truncation Approach for Various Levels of ϵ. Values based on the true data (dashed lines) and the prior information (dotted lines) are provided for reference, while the shaded bounds represent the 95 percent sampling interval of the synthetic data.

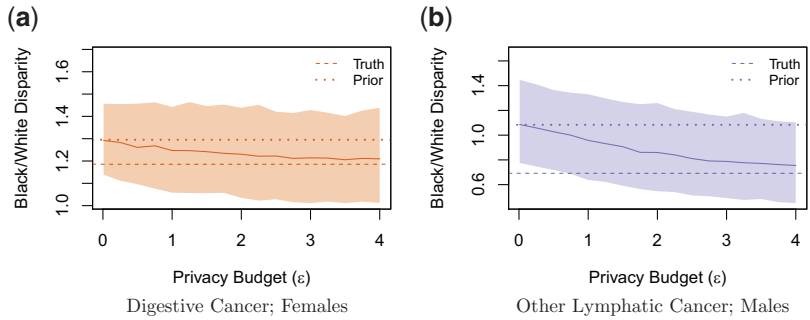

Figure 6. Estimated Black/White Disparities in Rates of Death due to Digestive Cancer (ICD-9: 150–159) among Women and in Rates of Death due to Other Lymphatic Cancer (ICD-9: 200–203) among Men Based on the Synthetic Data Across Various Levels of ϵ. Values based on the true data (dashed lines) and the prior information (dotted lines) are provided for reference, while the shaded bounds represent the 95 percent sampling interval of the synthetic data. (a) Digestive cancer; females.

result which is contrary to that at the national level. Due to the relatively low rate of death from other lymphatic cancer, obtaining evidence of a "statistically significant" black/white disparity from the synthetic data—in either direction—may be challenging for small values of ϵ, but merely obtaining estimates *suggesting* lower rates among white men may warrant further study (e.g., by analyzing the true data in a controlled environment).

While most of the results we have discussed thus far have been based on age-adjusted death rates for all types of cancer combined, it should be emphasized that the synthetic data in question were generated for each combination of age, race, sex, and specific type of cancer for each county. To this end, additional results such as investigations of geographic, urban/rural, and racial disparities in death rates for specific forms of cancer are provided in figures E.6–E.8 of the supplementary materials file online. Results for these subgroup analyses behave similar to the analyses presented in figures 3–6 in that when ϵ is large, the synthetic data essentially provide the same inference (on average) as the true data, and as ϵ decreases, inference based on the synthetic data slowly begin to drift toward the values from the prior distribution.

4. DISCUSSION

Quick (2021) generalized the important work of Machanavajjhala et al. (2008) for generating differentially private synthetic event counts from a multinomial-Dirichlet framework to a Poisson-gamma model, and in doing so allowed users to incorporate public knowledge about heterogeneity in population sizes and underlying event rates to improve the utility of the synthetic data. One limitation of the approach of Quick (2021) was that it was designed to protect against unrealistic worst-case scenarios—for example, the possibility of allocating all of our synthetic events to groups in which few events truly occurred. To address this issue, we proposed the use of prior predictive truncation to restrict the range of values that the synthetic counts can take based on their prior predictive expected values. As a result, the gamma priors used in this framework are significantly less informative than those proposed by Quick (2021), yielding a posterior distribution which leans more heavily on the true data and thus produces synthetic data with greater utility.

To address this issue in their own work, Machanavajjhala et al. proposed an (ϵ, δ)-probabilistically differentially private approach in which the synthetic data generated from their model would satisfy ϵ-differential privacy with probability $1 - \delta$. As with the approach proposed here, this allowed their approach to rely on less informative Dirichlet priors and thus improved the utility of the synthetic data. While the proposed approach and the framework of Machanavajjhala et al. (2008) are similar, there is one key distinction between the two approaches: because the approach of Machanavajjhala et al. depends on the probability of sampling values of $\mathbf{z}$ that violate ϵ-differential privacy, the hyperparameters underlying their Dirichlet priors, $\boldsymbol{\alpha}$, are a function of the true data and thus cannot be publicly disclosed without further violations of differential privacy. In contrast, the restrictions imposed on $\mathbf{z}$ described in section 2.3 are—by definition—based on the prior predictive distribution and thus are based solely on *publicly available* information. As a result, the restrictions imposed on $\mathbf{z}$ and the values of the hyperparameters, $\mathbf{a}$ and $\mathbf{b}$, can be released

without leaking sensitive information about the true data. Previous work (e.g., Charest, 2011) has considered treating synthetic data as noisy versions of the truth and using measurement error models (informed by the true hyperparameters) in an attempt to remove the differentially private noise and recover the true data—while unexplored here, disclosing hyperparameters like **a** and **b** should help facilitate analyses of this nature.

While the proposed approach is capable of producing differentially private synthetic data with high utility, there is one key caveat—the utility of the synthetic data under this approach (and its predecessor, the approach of Quick, 2021) strongly depends on the quality of the prior information. As a result, the use of the Poisson-gamma model with prior predictive truncation proposed here should only be considered in settings where high-quality prior information is available and can be used with confidence. For instance, Quick (2021) demonstrated that the multinomial-Dirichlet framework of Machanavajjhala et al. (2008) can perform poorly in applications like CDC WONDER where a high degree of heterogeneity in population sizes and underlying event rates can occur, as these properties are in conflict with the framework's (implicit) assumption of homogeneity in the prior expected values. Similarly, the prior information used in section 3 assumes that prior event rates are uniform across geographic areas—while an assumption of this nature may suffice for synthesizing deaths due to *chronic* diseases like heart disease and cancer, accounting for disparities between urban and rural regions may be crucial for other causes of death such as drug overdose (Paulozzi and Xi, 2008) and infectious diseases (e.g., Zhang and Schwartz, 2020). In these scenarios—for example, when events are expected to cluster geographically but knowledge about *where* these clusters are is not known (or otherwise unaccounted for)—the bounds in (9) can be relaxed by the inclusion of an inflation/deflation factor, $\xi \geq 1$, such that

$$L_i = F^{-1}(\alpha/2|\xi^{-1}n_i\lambda_{i0}) \leq z_i \leq F^{-1}(1-\alpha/2|\xi n_i\lambda_{i0}) = U_i.$$

For instance, Paulozzi and Xi (2008) reported that rates of unintentional and undetermined intent drug poisoning death related to heroin/opium were >5 times higher in large central metro counties (0.93 deaths per 100,000) than in noncore (rural) counties (0.18) in 2004; in that setting, a value of $\xi > 2$ might be required to construct bounds that adequately cover the range of death rates across the spectrum of urbanization. In short, the approach proposed here is *not* intended for use as a black-box algorithm but rather as an approach to be used in a very deliberate manner. To this end, we believe agencies interested in using this approach should first identify historical data that can be used as a testbed to calibrate the framework for future data releases—for example, *What sources of prior information are available?* and *What "known" features are unaccounted for in the prior information?*

Another important topic for discussion is that of the *availability* of external information. In analysis of the Pennsylvania cancer data, we utilized two

sources of external information: population estimates provided (in part) by the Census Bureau and national death rates from the CDC's annual death reports. While it is true that future releases of the bridged-race population estimates produced by the CDC and the Census Bureau will be protected using differential privacy, early indications from the 2020 Decennial Census of Population and Housing (e.g., Wright and Irimata, 2021) are such that the accuracy of public-use data will remain high, at least at the level of granularity necessary for applications like CDC WONDER. Thus, the greater concern with respect to data access in the context of a Synthetic CDC WONDER is the continued availability of national estimates of death rates published in the CDC's annual death reports. Here, the CDC would seemingly have two options. On one hand, they could simply decide to release certain national estimates with *no* privacy protections. While this would, strictly speaking, violate differential privacy, holding certain quantities invariant would be consistent with the Census Bureau's implementation of the TopDown algorithm (Abowd et al., 2019). Alternatively, it is also conceivable that the CDC would apply a differentially private mechanism—for example, adding Laplace noise—to the true national death counts prior to their inclusion in the annual death reports. In this scenario, one might view the *true* privacy budget for the Synthetic CDC WONDER, ϵ, as being partitioned into a privacy budget for the *county-level* data, ϵ_{County}, and a privacy budget for the *national* data, ϵ_{Nat}, such that $\epsilon = \epsilon_{\text{County}} + \epsilon_{\text{Nat}}$. Fortunately, because the associated denominators at the national level—that is, the total number of people by age, race, and sex in the United States—should all be quite large, noise added to the death counts should have little impact on the accuracy of the death rates for even small levels of ϵ_{Nat}.

Finally, we would be remiss to not acknowledge that conventional approaches for satisfying differential privacy are not based on posterior predictive synthesis as described here. For instance, approaches such as the Laplace mechanism (Dwork et al., 2006) and the geometric mechanism (Ghosh et al., 2012) operate by adding noise to the true counts. While the objective of this paper was to demonstrate the improvement in utility associated with prior predictive truncation, comparisons between the Poisson-gamma framework of Quick (2021)—with or without truncation—and existing, more conventional approaches for satisfying differential privacy must be conducted before it should be implemented in practice. It should also be noted that both in this paper and in the work of Quick (2021), the "utility" of the synthetic data was primarily assessed as a function of *bias*—for example, as ϵ decreases, the model puts more emphasis on the prior and estimates based on the synthetic data drift away from the true values—while the *variability* associated with the synthetic data was *shown* but not a *focal point* in the evaluation of the synthetic data. In the context of the Poisson-gamma framework, this focus on bias over variability is justified as the variance of the synthetic data is largely driven by the likelihood rather than the prior (e.g., the variance of the sampling distributions in

figures 5 and E.2 are similar for all values of ϵ). In these broader comparisons, however, a more balanced assessment of the bias and the variability of estimates produced by these approaches will be necessary to fully appreciate their respective strengths and weaknesses.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

Abowd, J., R. Ashmead, G. Simson, D. Kifer, P. Leclerc, A. Machanavajjhala, and W. Sexton (2019), "Census TopDown: Differentially Private Data, Incremental Schemas, and Consistency with Public Knowledge," Technical Report. US Census Bureau.

Abowd, J. M. (2018), "The U.S. Census Bureau Adopts Differential Privacy," Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '18, 2867, New York, NY, USA: ACM.

Besag, J., J. York, and A. Mollié (1991), "Bayesian Image Restoration, with Two Applications in Spatial Statistics," *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Brillinger, D. R. (1986), "The Natural Variability of Vital Rates and Associated Statistics," *Biometrics*, 42, 693–734.

CDC. (2003), "CDC/ATSDR Policy on Releasing and Sharing Data." Manual; Guide CDC-02. Available at http://www.cdc.gov/maso/Policy/ReleasingData.pdf. Accessed 24 February 2022.

Charest, A.-S. (2011), "How Can we Analyze Differentially-Private Synthetic Datasets?," *Journal of Privacy and Confidentiality*, 2, 21–33.

Dinur, I., and K. Nissim (2003), "Revealing Information While Preserving Privacy," Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '03, pp. 202–210. New York, NY, USA: ACM.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006), "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of Cryptography*, eds. S. Halevi and T. Rabin, pp. 265–284. Berlin/Heidelberg: Springer.

Friede, A., J. A. Reid, and H. W. Ory (1993), "CDC WONDER: A Comprehensive on-Line Public Health Information System of the Centers for Disease Control and Prevention," *American Journal of Public Health*, 83, 1289–1294.

Gelfand, A. E., and P. Vounatsou (2003), "Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis," *Biostatistics*, 4, 11–25.

Ghosh, A., T. Roughgarden, and M. Sundararajan (2012), "Universally Utility-Maximizing Privacy Mechanisms," *SIAM Journal on Computing*, 41, 1673–1693.

Holan, S. H., D. Toth, M. A. R. Ferreira, and A. Karr (2010), "Bayesian Multiscale Multiple Imputation with Implications for Data Confidentiality," *Journal of the American Statistical Association*, 105, 564–577.

Ingram, D. D., J. D. Parker, N. Schenker, J. A. Weed, B. Hamilton, E. Arias, and J. H. Madans (2003), "United States Census 2000 Population with Bridged Race Categories," National Center for Health Statistics. Vital and Health Statistics Reports, Series, 2.

Kerman, J. (2011), "Neutral Noninformative and Informative Conjugate Beta and Gamma Prior Distributions," *Electronic Journal of Statistics*, 5, 1450–1470.

Kochanek, K. D., S. L. Murphy, J. Xu, and E. Arias (2019), "Deaths: Final Data for 2017," *National Vital Statistics Reports*, 68, 1–77.

Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber (2008), "Privacy: Theory Meets Practice on the Map," IEEE 24th International Conference on Data Engineering, pp. 277–286.

Murphy, S. L., J. Xu, K. D. Kochanek, E. Arias, and B. Tejada-Vera (2021), "Deaths: Final Data for 2018," *National Vital Statistics Reports*, 69, 1–83.

National Center for Health Statistics. (1985), "Vital Statistics of the United States, 1980. Volume II, Mortality, Part A," DHHS Pub. No. (PHS) 85-1101, Public Health Service. Washington: U.S. Government Printing Office.

Paulozzi, L. J., and Y. Xi (2008), "Recent Changes in Drug Poisoning Mortality in the United States by Urban-Rural Status and by Drug Type," *Pharmacoepidemiology and Drug Safety*, 17, 997–1005.

Quick, H. (2021), "Generating Poisson-Distributed Differentially Private Synthetic Data," *Journal of the Royal Statistical Society, Series A*, 184, 1093–1108.

Quick, H., S. H. Holan, and C. K. Wikle (2015), "Zeros and Ones: A Case for Suppressing Zeros in Sensitive Count Data with an Application to Stroke Mortality," *Stat*, 4, 227–234.

Quick, H., and L. A. Waller (2018), "Using Spatiotemporal Models to Generate Synthetic Data for Public Use," *Spatial and Spatio-Temporal Epidemiology*, 27, 37–45.

Wright, T., and K. Irimata (2021), "Empirical Study of Two Aspects of the TopDown Algorithm for Redistricting: Reliability and Variability," Technical Report. US Census Bureau.

Zhang, C. H., and G. G. Schwartz (2020), "Spatial Disparities in Coronavirus Incidence and Mortality in the United States: An Ecological Analysis as of May 2020," *Journal of Rural Health*, 36, 433–445.