Improving the Utility of Poisson-Distributed, Differentially Private Synthetic Data via Prior Predictive Truncation with an Application to CDC WONDER

Harrison Quick (Drexel University)

Funding for this research came from the National Science Foundation (NSF-SES-1943730)

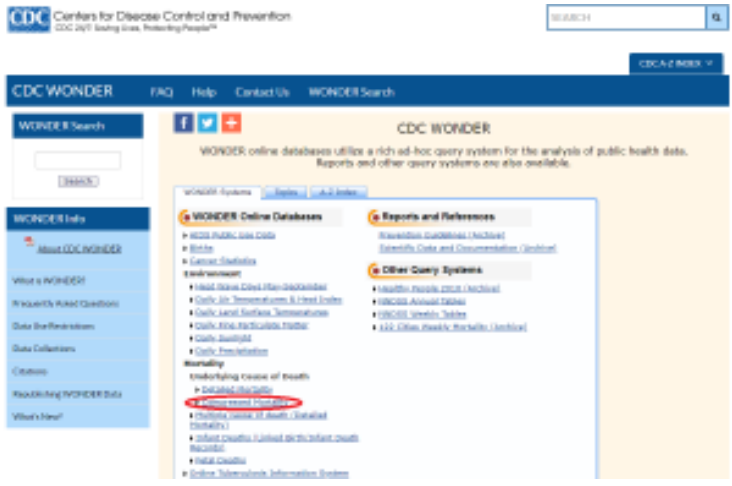
Table of Contents

Table of Contents

CDC WONDER


Centers for Disease Control and Prevention
CDC WONDER
FAQ
Help
Contact Us
WONDER Search
WONDER Search
CDC WONDER
WONDER online databases utilize a rich ad-hoc query system for the analysis of public health data.
Reports and other query systems are also available.
WONDER Online Databases
Reports and References
Other Query Systems

CDC WONDER


Centers for Disease Control and Prevention
CDC WONDER
FAQ
Help
Contact Us
WONDER Search
WONDER Search
CDC WONDER
WONDER online databases utilize a rich ad-hoc query system for the analysis of public health data.
Reports and other query systems are also available.
WONDER Online Databases
Reports and References
Other Query Systems

CDC WONDER

For the sake of illustration, we'll look at county-level data for ages 35–44 from 2016

CDC WONDER

These ICD-10 codes represent “death due to heart disease”

CDC WONDER

County-level heart disease-related death counts for ages 35–44 in 2016 from all races and all genders

All counts less than 10 are suppressed in public-use datasets

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
  - Differences by age
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
  - Differences by age
  - Differences by cause-of-death
- Privacy

Is there a way that CDC can address these issues?

CDC WONDER

While CDC WONDER offers a wealth of data and *does* implement privacy protections, there is still room for improvement:

- Utility: Suppression of small counts affects users' ability to assess...
  - Urban/Rural disparities
  - Racial disparities
  - Differences by sex
  - Differences by age
  - Differences by cause-of-death
- Privacy
  - Targeted attacks by clever intruders can overcome data suppression to uncover the true counts

Is there a way that CDC can address these issues?

Synthetic Data

One option to address the issue of data suppression would be to release *synthetic data*: e.g., if

- $\mathbf{y} = (y_1, \ldots, y_l)^T$ denotes a restricted-use dataset,
- $p(\mathbf{y} \mid \phi)$ is an appropriate statistical model for $\mathbf{y}$ with parameters ϕ, and
- $p(\phi \mid \psi)$ is a prior distribution for ϕ given hyperparameters, ψ,

then we can generate a synthetic dataset, $\mathbf{z} = (z_1, \ldots, z_l)^T$, from the posterior predictive distribution,

$$p(\mathbf{z} \mid \mathbf{y}, \psi) = \int p(\mathbf{z} \mid \phi)\, p(\phi \mid \mathbf{y}, \psi)\, d\phi.$$

More specifically, we can sample ϕ^* from $p(\phi \mid \mathbf{y}, \psi)$ and then sample $\mathbf{z}$ from $p(\mathbf{z} \mid \phi^*)$.

Differentially Private Synthetic Data(Dwork, 2006)

The standard typically used for demonstrating formal privacy guarantees is the concept of *differential privacy* (Dwork, 2006).

In this context, $p(\mathbf{z} \mid \mathbf{y}, \psi)$ is ϵ-differentially private if for any similar[1] dataset, $\mathbf{x}$,

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

While ψ *can* be viewed as a vector of model parameters — selected *a priori* in hopes that $E[y_i \mid \psi] \approx y_i$ in order to produce synthetic data with high utility — the elements of ψ are *primarily* used to satisfy ϵ-differential privacy.

[1] $\|\mathbf{x} - \mathbf{y}\| = 2$ and $\sum_i x_i = \sum_i y_i$ — i.e., there exists i and i' such that $x_i = y_i - 1$ and $x_{i'} = y_{i'} + 1$ with all other values equal

Table of Contents

Multinomial-Dirichlet model

Let $\mathbf{y}$ be a vector of sensitive count data of length $I \geq 2$ with $\sum_i y_i = y_.$ and assume

$$\mathbf{y} \,|\, \boldsymbol{\theta} \sim \text{Mult}\,(y_., \boldsymbol{\theta}) \text{ and } \boldsymbol{\theta} \sim \text{Dir}\,(\boldsymbol{\alpha})\,.$$

It can (but won't) be shown that if

$$\min \alpha_i \geq z_. / \left[\exp\left(\epsilon\right) - 1\right],$$

the multinomial-Dirichlet synthesizer, $p\left(\mathbf{z} \,|\, \mathbf{y}, \boldsymbol{\alpha}\right)$, will satisfy ϵ-differential privacy.

- i.e., if our $\text{Dir}\,(\boldsymbol{\alpha})$ prior is informative enough, it can sufficiently mask the data.

Key drawback: Assumes *homogeneity*

- Shouldn't Philadelphia have more deaths than Small Town, PA?
- Shouldn't more deaths be attributed to old people than young people?

Poisson-Gamma model

In contrast, the Poisson-gamma framework assumes

$$y_i \mid \lambda_i \sim \text{Pois}\left(n_i \lambda_i\right) \text{ and } \lambda_i \sim \text{Gamma}\left(a_i, b_i\right).$$

Since the y_i are (conditionally) independent Poisson random variables, we can write

$$\mathbf{y} \mid \boldsymbol{\lambda}, \sum_i y_i = y_\cdot \sim \text{Mult}\left(y_\cdot, \left\{ \frac{n_i \lambda_i}{\sum_j n_j \lambda_j} \right\}\right)$$

- Allows for heterogeneity in population sizes (via n_i) and underlying event rates (via a_i / b_i)

But under what conditions will this satisfy ϵ-differential privacy?

Poisson-Gamma model

It *can* (but won't) be shown that the Poisson-gamma synthesizer, denoted $p(\mathbf{z} \mid \mathbf{y}, \mathbf{a}, \mathbf{b})$, will satisfy ϵ-differential privacy if

$$a_i \geq \frac{z_.}{e^{\epsilon}/\nu_i - 1} \tag{2}$$

where $\nu_i \in [1, 2]$ denotes what amounts to a *penalty* term associated with the additional information gained from using the Poisson-gamma model compared to the multinomial-Dirichlet model.

Key drawback: Extreme "worst case scenario"

- Above criteria protects against group with ONE observed event ($y_i = 1$) being assigned ALL of the synthetic events ($z_i = z_.$).
- e.g., all cancer-related deaths in PA being assigned to a single rural county — this *shouldn't* be possible, so why should we worry about this???

Prior predictive truncated Poisson-gamma framework

Rather than focus on technical details, let's consider a hypothetical example.

Suppose $E[y_i \mid \mathbf{a}, \mathbf{b}, \mathbf{n}] = n_i \lambda_{i0} = 10$ for a given i and that our dataset consists of $y_{\cdot} > 26{,}000$ events. Then 99.9% of the prior predictive distribution falls between $y_i = 2$ and $y_i = 22$.

```
> qpois(.0005,10)
[1] 2
> qpois(.9995,10)
[1] 22
```

- If $y_{\cdot} > 26{,}000$, we should expect a reduction in model informativeness on the order of

$$\frac{26{,}000}{22 - 2} > 1{,}300$$

Note: This approach is *heavily* dependent on having high quality prior information

- If $E[y_i \mid \mathbf{a}, \mathbf{b}, \mathbf{n}] = n_i \lambda_{i0} \not\approx y_i$, then the prior predictive bounds will not be good.
- We will need to rely on subject-matter experts to know what is sufficient

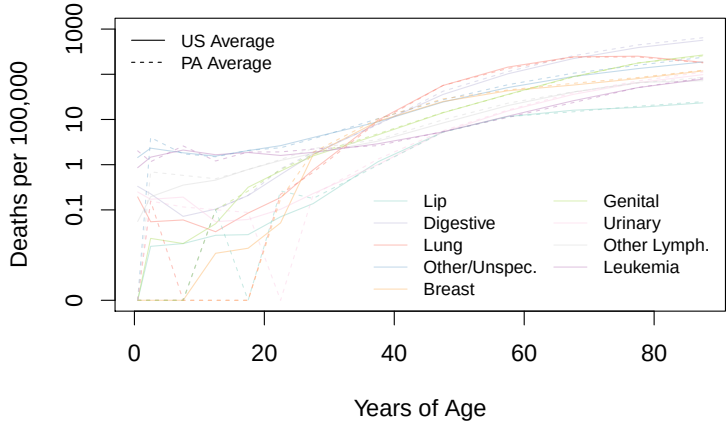
Table of Contents

PA cancer death data from 1980 — 26,116 deaths from 47,034 strata

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 under 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 Other)
Sex	$s = 1, 2$ Levels (Male and Female)

Overview of the structure of the Pennsylvania cancer data. Cancer types are identified by their International Classification of Diseases, Ninth Revision (ICD-9) codes. Data are publicly available — free of suppression — because they predate the privacy protections.

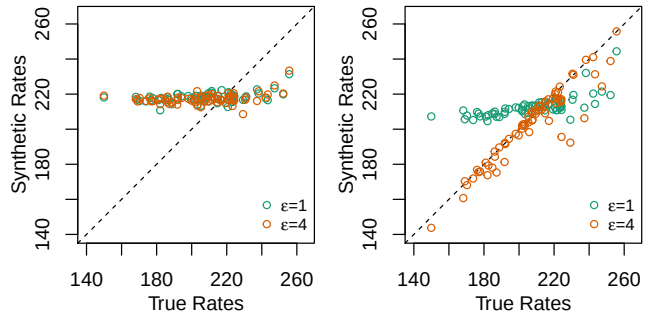
Prior information: National death rates vs. PA death rates



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.

- We don't need these to be *perfect*, we just need them to *comparable*.

Utility of synthetic data: Age-adjusted rates

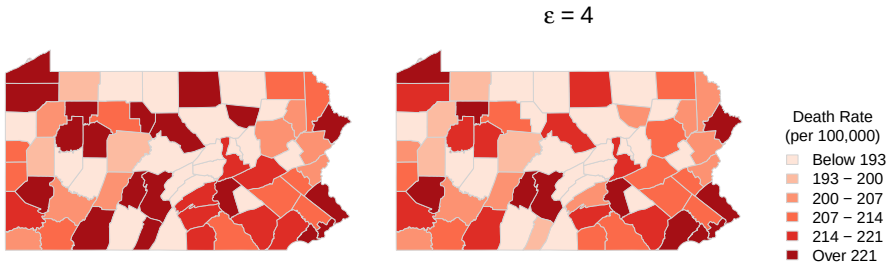


(a) Original/Untruncated (b) Prior Predictive Truncation

Figure 1: Comparison of age-adjusted cancer-related death rates based on the two approaches for generating synthetic data for $\epsilon = 1$ and $\epsilon = 4$.

- When $\epsilon = 1$, the original model requires all $a_i > 15{,}000$, whereas the prior predictive truncation approach has a $\max(a_i) < 17$ and most are less than 0.58

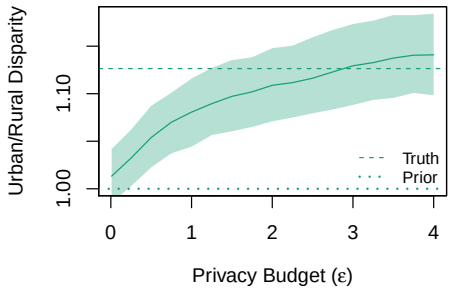
Utility of synthetic data: Age-adjusted rates



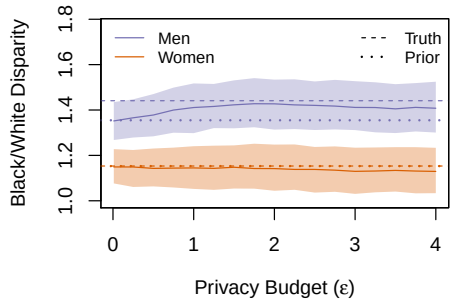
(a) True Age-Adjusted Rates (b) Synthetic Age-Adjusted Rates

- Because the prior information does not account for *geographic* disparities, as $\epsilon \to 0$, estimates become geographically homogenous
- Point of emphasis: More difficult to identify *true* disparities, but also unlikely to produce *spurious* disparities

Utility of synthetic data: Urban/rural and black/white disparities



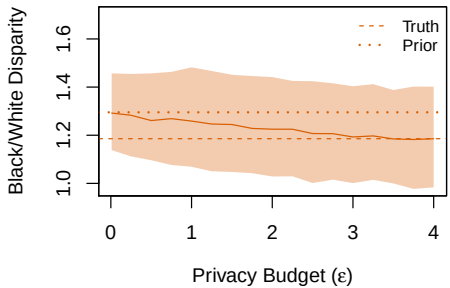
(a) Urban/Rural Disparity



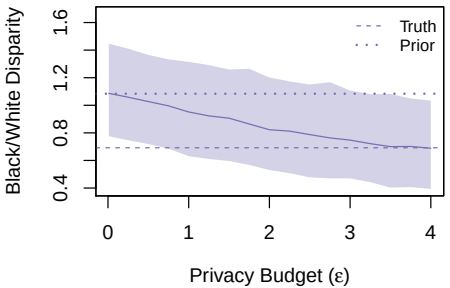
(b) Black/White Disparity

- Here again, as $\epsilon \to 0$, our estimates of the disparities shift from "the truth" to "the prior".
  - Prior information did not include anything about urban/rural disparities, so the effect is attenuated toward the null (i.e., no disparity)

Utility of synthetic data: Why we need to release the prior information



(a) Digestive Cancer; Females



(b) Other Lymphatic Cancer; Males

- Digestive Cancer; Females: National black/white disparity is *larger* than in PA
- Other Lymphatic Cancer; Males: National black/white disparity is *the opposite* of the disparity in PA
  - Disclosing the prior information will help users determine if the results from the synthetic data are driven by the data or are a reflection of the prior

Table of Contents

Summary

Using the prior predictive distribution to truncate the range of values for the Poisson-gamma model can reduce the model's informativeness by several orders of magnitude, thereby producing *substantial* increases in utility

- The utility of the prior predictive truncation approach is heavily reliant on the quality of the prior information; e.g., mortality rates differ by age, thus our prior information ought to differ by age
- A small amount of our privacy budget can be used to protect the prior information. Ideally, the prior would be based on relatively large counts (e.g., national death counts) such that adding DP noise would be unlikely to cause any meaningful changes.

Thanks for listening!

References:

- Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). "Privacy: Theory meets practice on the map." In *IEEE 24th International Conference on Data Engineering*, 277–286.
- Quick, H. (2021). "Generating Poisson-distributed differentially private synthetic data." *J. Roy. Statist. Soc., Ser. A (Statistics in Society)*, **184**, 1093–1108.
- Quick, H. "Improving the utility of Poisson-distributed, differentially private synthetic data via prior predictive truncation with an application to CDC WONDER." arXiv preprint 2103.03833.