

The Intersection of Informative Priors and Differential Privacy in Bayesian Spatial Biostatistics

Harrison Quick (University of Minnesota)

Table of Contents

Motivating Use-Case: CDC WONDER

Cancer-related Deaths in Pennsylvania Counties in 1980

Summary & Discussion

Table of Contents

Motivating Use-Case: CDC WONDER

Cancer-related Deaths in Pennsylvania Counties in 1980

Summary & Discussion

WONDER Search

WONDER Info

[About CDC WONDER](#)[What is WONDER?](#)[Frequently Asked Questions](#)[Data Use Restrictions](#)[Data Collections](#)[Citations](#)[Republishing WONDER Data](#)[What's New?](#)

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 Systems](#) [Topics](#) [A-Z Index](#)

WONDER Online Databases

- ▶ [AIDS Public Use Data](#)
- ▶ [Births](#)
- ▶ [Cancer Statistics](#)
- Environment**
 - ▶ [Heat Wave Days May-September](#)
 - ▶ [Daily Air Temperatures & Heat Index](#)
 - ▶ [Daily Land Surface Temperatures](#)
 - ▶ [Daily Fine Particulate Matter](#)
 - ▶ [Daily Sunlight](#)
 - ▶ [Daily Precipitation](#)
- Mortality**
 - Underlying Cause of Death**
 - ▶ [Detailed Mortality](#)
 - ▶ [Compressed Mortality](#)
 - ▶ [Multiple cause of death \(Detailed Mortality\)](#)
 - ▶ [Infant Deaths \(Linked Birth/Infant Death Records\)](#)
 - ▶ [Fetal Deaths](#)
 - ▶ [Online Tuberculosis Information System](#)

Reports and References

- ▶ [Prevention Guidelines \(Archive\)](#)
- ▶ [Scientific Data and Documentation \(Archive\)](#)

Other Query Systems

- ▶ [Healthy People 2010 \(Archive\)](#)
- ▶ [NNDSS Annual Tables](#)
- ▶ [NNDSS Weekly Tables](#)
- ▶ [122 Cities Weekly Mortality \(Archive\)](#)

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

Compressed Mortality, 1999-2016 Results

Request Form Results Map Chart About

[Compressed Mortality Data](#) [Dataset Documentation](#) [Other Data Access](#) [Help for Results](#) [Printing Tips](#) [Help with Exports](#)

[Notes](#) [Citation](#) [Query Criteria](#)

County ↓	Deaths ↑↓	Population ↑↓	Crude Rate Per 100,000 ↑↓
Autauga County, AL (01001)	Suppressed	7,190	Suppressed
Baldwin County, AL (01003)	14	24,545	57.0 (Unreliable)
Barbour County, AL (01005)	Suppressed	3,171	Suppressed
Bibb County, AL (01007)	Suppressed	3,043	Suppressed
Blount County, AL (01009)	Suppressed	7,090	Suppressed
Bullock County, AL (01011)	Suppressed	1,301	Suppressed
Butler County, AL (01013)	Suppressed	2,262	Suppressed
Calhoun County, AL (01015)	19	13,460	141.2 (Unreliable)

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

- ▶ $\mathbf{y} = (y_1, \dots, y_I)^T$ denotes a restricted-use dataset of I observations,
- ▶ $p(\mathbf{y} | \phi)$ is an appropriate statistical model for \mathbf{y} with parameters ϕ , and
- ▶ $p(\phi | \psi)$ is a prior distribution for ϕ given hyperparameters, ψ ,

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

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

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

- ▶ Natural next question: How do we know if synthetic data generated from $p(\mathbf{z} | \mathbf{y}, \psi)$ are sufficiently protective?

Differential Privacy (Dwork et al., 2006)

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

In this context, $p(\mathbf{z} | \mathbf{y}, \psi)$ is ϵ -differentially private if for any similar¹ dataset, \mathbf{x} ,

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

While ψ can be viewed as a vector of model parameters, *in practice* the elements of ψ are merely specified to satisfy ϵ -differential privacy.

- ▶ What are some options for $p(\mathbf{y} | \phi)$ and $p(\phi | \psi)$ that yield a $p(\mathbf{z} | \mathbf{y}, \psi)$ that can be shown to satisfy 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

What is Differential Privacy? A Simple (Conventional) Example...

True Data		Intruder's Data	
△		?	
△		△	
△	+	△	+
△	+	△	+
△	+	△	+
△	+	△	+
△	+	△	+
△	+	△	+

Suppose we want to release the proportion of triangles in this dataset without disclosing any individual shape.

- ▶ Worst case scenario: Intruder knows all but one shape
 - ▶ Releasing the true value (8/14) compromises the remaining shape
- ▶ Let's add noise to the proportion such that

$$\frac{8 + \text{noise}}{14} \approx \frac{7 + \text{noise}}{14}$$

where the amount of noise depends on the level of protection desired (measured by ϵ)

- ▶ e.g., noise \sim Lap(0, 1/ ϵ)

A Working Hypothesis...

The **hypothesis** underlying my work in data privacy is that **synthetic data generated from the true data generating process** will **outperform a noisy version of the true data**.

- ▶ For instance, approaches that merely add noise operate on an *absolute* scale, whereas many public health outcomes operate on *relative* scales.
 - ▶ e.g., 200 deaths vs. 210 deaths compared to 0 deaths vs. 10 deaths.

Thus, my strategy is to specify a model that (a) is statistically appropriate and (b) can be proven to satisfy differential privacy, and then hope the posterior predictive distribution is a good approximation of the true data generating process.

- ▶ In the next few slides, I'll discuss two simple model specifications — the multinomial-Dirichlet model and the Poisson-gamma model — that have been proven to satisfy differential privacy.
- ▶ Along the way, I will discuss the appropriateness of these models for synthesizing public health data — e.g., county-level death counts.

Multinomial-Dirichlet Model (Machanavajjhala et al., 2008)

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

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

To generate a synthetic data vector, \mathbf{z} , with a given $\sum_i z_i = z = y$:

1. Sample $\boldsymbol{\theta}^*$ from its posterior, $\boldsymbol{\theta} | \mathbf{y} \sim \text{Dir}(\mathbf{y} + \boldsymbol{\alpha})$
2. Sample \mathbf{z} from the posterior predictive distribution, $\mathbf{z} \sim \text{Mult}(z, \boldsymbol{\theta}^*)$

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

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

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

- ▶ If our $\text{Dir}(\boldsymbol{\alpha})$ prior is **informative** enough, it can sufficiently mask the data...
 - ▶ ... but it will do so by allocating events *uniformly*, which is bad.
 - ▶ e.g., if ϵ is small, the model will try to assign the same number of events to Small Town, PA as it would to Philadelphia.

Poisson-Gamma Model (Quick, 2021)

Motivated by the field of disease mapping — where event counts are typically modeled as being Poisson distributed — Quick (2021) proposed assuming

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

which implies $\lambda_i | y_i \sim \text{Gamma}(y_i + a_i, n_i + b_i)$. Now recall that if the y_i are (conditionally) independent Poisson random variables, then

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

Thus, we can generate synthetic data by:

1. Sampling λ_i^* from $\text{Gamma}(y_i + a_i, n_i + b_i)$ for $i = 1, \dots, l$
2. Sampling $\mathbf{z} \sim \text{Mult} \left(z., \left\{ n_i \lambda_i^* / \sum_j n_j \lambda_j^* \right\} \right)$

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

Poisson-Gamma Model — ϵ -differential privacy

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

$$a_i \geq \frac{z_i}{e^\epsilon / \nu_i - 1} \quad (2)$$

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.

- ▶ It would take too much time/space to write out the expression for ν_i , but it's a function of the group-specific population sizes and prior event rates.
- ▶ If the group-specific population sizes and prior event rates are equal, then $\nu_i = 1$ for all groups, thus making the M-D and P-G models *mathematically equivalent*.

Drawback of the Poisson-Gamma Model of Quick (2021)

Unlike the multinomial-Dirichlet model, the Poisson-gamma model behaves fairly well when ϵ is small.

- ▶ i.e., the model will allocate events based on the population sizes, n_i , and the prior expected event rates, $\lambda_{i0} = a_i/b_i$, thus if these values were chosen “wisely”, we won’t get *terrible* synthetic data like Small Town, PA \approx Philadelphia

Unfortunately, another problem with the multinomial-Dirichlet model that *is* shared by the Poisson-gamma model of Quick (2021) is that when the total number of events, y ., is large, *very* informative priors are required to satisfy only moderate values of ϵ .

- ▶ While it is unlikely that Small Town, PA would be assigned as many events as Philadelphia, our privacy protections are designed to guard against this possibility, and that’s where the issues arise.

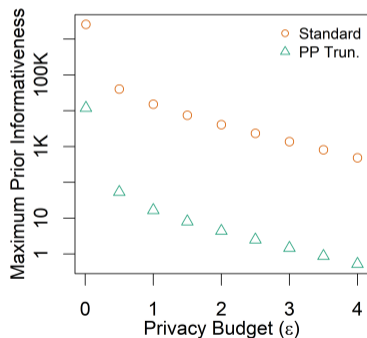
As a result, the synthetic data typically just reflect the prior information.

Prior Predictive Truncated Poisson-Gamma Model (Quick, 2022)

To combat this, Quick (2022) proposed using the *prior predictive distribution* to truncate the synthetic data to a “reasonable” range of values.

- ▶ e.g., if $E[y | n, a, b] = 10$ deaths, we can confidently assume that $y < 30$.

This allows us to use *far less informative priors* which in turn yields synthetic data with *far greater utility*.



Visualizing model informativeness and differential privacy...

Consider a scenario in which $y_i = 5$ and $n_i = 625$ and a prior $\lambda_i \sim \text{Gamma}(a, b)$ whose mean corresponds to 300 deaths per 100,000. In a differentially private framework, we need to guard against the scenario in which a hypothetical intruder's data is $x_i = y_i - 1 = 4$ and $n_i = 625$.

- ▶ When using noninformative priors, the difference between the posterior distributions for λ_i is pretty clear
 - ▶ Good for utility, bad for privacy
- ▶ As a increases, both posteriors get pulled toward λ_0 and the differences between them become more subtle
 - ▶ Bad for utility, good for privacy

Note: *Technically*, we need to compare the posterior predictive distributions — $p(\mathbf{z} | \mathbf{y}, \mathbf{a}, \mathbf{b})$ and $p(\mathbf{z} | \mathbf{x}, \mathbf{a}, \mathbf{b})$ — but it's the same concept.

Table of Contents

Motivating Use-Case: CDC WONDER

Cancer-related Deaths in Pennsylvania Counties in 1980

Summary & Discussion

Cancer-related Deaths in Pennsylvania Counties in 1980

Attribute	Levels
County	$i = 1, \dots, 67$ Counties in Pennsylvania
Cancer Type	$c = 1, \dots, 9$ 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	$a = 1, \dots, 13$ 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	$r = 1, \dots, 3$ Levels (Black, White, and Other)
Sex	$s = 1, 2$ Levels (Male and Female)

In total, there were $y. = \sum_{icars} y_{icars} = 26,116$ cancer-related deaths in PA in 1980 belonging to these $67 \times 13 \times 9 \times 3 \times 2 = 47,034$ strata.

How good is our prior information?

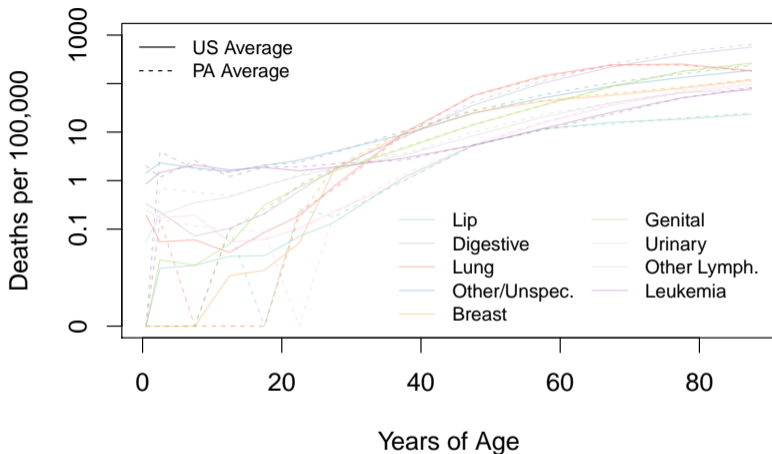


Figure 1: Cause-specific death rates at the national level and for the state of Pennsylvania. National-level rates are used as prior information for estimating the proper allocation of deaths at the state and county level.

What do the synthetic data look like?

(a) Group with small y , $E(y | \mathbf{a}, \mathbf{b})$

(b) Group with large y , $E(y | \mathbf{a}, \mathbf{b})$

Figure 2: Posterior predictive distribution for various levels of ϵ . In Panel (a), the prior predictive expected value is $E[y | \mathbf{a}, \mathbf{b}] = 1.15$ and the true death count is $y = 0$. In Panel (b), the prior predictive expected value is $E[y | \mathbf{a}, \mathbf{b}] = 211$ and the true death count is $y = 237$.

- ▶ As $\epsilon \rightarrow 0$, the synthetic values shift away from y toward $E[y | \mathbf{a}, \mathbf{b}]$.
- ▶ Moreover, note that the variability of the synthetic data differ — data protected by the Laplace mechanism would not have this feature.

Age-Adjusted Cancer Death Rates



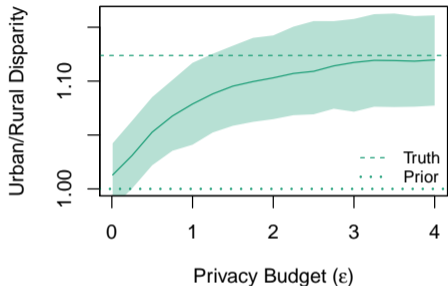
(a) True Age-Adjusted Rates

(b) Synthetic Age-Adjusted Rates

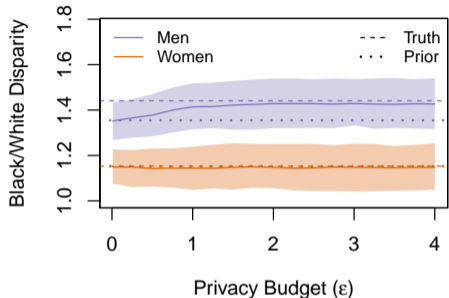
Figure 3: Degradation in utility for the age-adjusted rates as ϵ decreases.

- ▶ For large ϵ , geographic disparities in the data are largely preserved
- ▶ As $\epsilon \rightarrow 0$, the prior — which does not account for geographic disparities — becomes more influential and the rates all converge toward the statewide average

Disparities in Cancer Death Rates



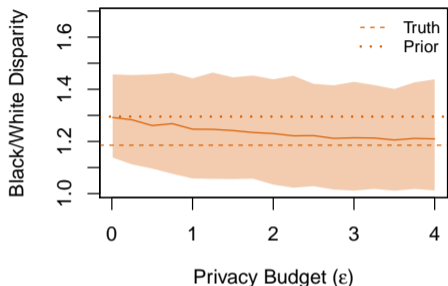
(a) Urban/Rural Disparity



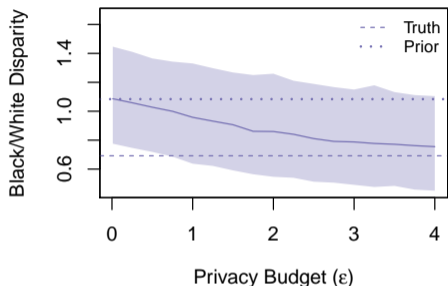
(b) Black/White Disparity

Figure 4: 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 variability of the synthetic data.

Disparities in Cause-Specific Cancer Death Rates



(a) Digestive Cancer; Females



(b) Other Lymphatic Cancer; Males

Figure 5: 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% sampling interval of the synthetic data.

Table of Contents

Motivating Use-Case: CDC WONDER

Cancer-related Deaths in Pennsylvania Counties in 1980

Summary & Discussion

Summary

- ▶ The CDC's current privacy protections — i.e., the suppression of small counts — are *ineffective* from a privacy standpoint and *discourage* research on important topics related to health disparities
- ▶ The release of differentially private synthetic data *would have* improved privacy protections and *could* encourage and facilitate research on health disparities
- ▶ The proposed Poisson-gamma framework is *promising*, but the utility of the synthetic data it produces is highly dependent on the prior information
 - ▶ My student and I are working on extending the methods here from the simple Poisson-gamma framework to a framework that is more consistent with the spatial statistics / disease mapping literature. This should be less reliant on “prior information” and should yield synthetic data with higher utility for small levels of ϵ .

Discussion

- ▶ The notion of using Bayesian models with informative priors to produce synthetic data with strong privacy protections should extend beyond Poisson-distributed count data, but *proving* that a given model specification can satisfy differential privacy could be very challenging.
- ▶ More broadly speaking, the privacy community is split between those who believe that having *provable* privacy protections is a necessity and those who believe that this is too strict / impractical.
 - ▶ I'm somewhere in between: If a differentially private method satisfies your needs, that's great. If not, a carefully constructed framework for generating synthetic data can still strike a balance between utility and privacy.

Concluding Remarks

- ▶ Working with (and *having access to*) rich public health data is crucial for disease surveillance and epidemiologic research.
- ▶ While there are valid privacy concerns associated with releasing vital statistics data, there *are* ways to disseminate rich public health data with high utility and strong privacy protections.
- ▶ That said, we still have some work ahead to dispel myths about the nature of synthetic data to the broader public health community...

Acknowledgements and References

Acknowledgements

- ▶ CDC/DHDSP for years of collaboration
- ▶ NSF, NCHS, the County Health Rankings & Roadmaps program, and the UMN's Data Science Initiative for funding my privacy research

References:

- ▶ Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography*, eds. S. Halevi and T. Rabin, 265–284. Berlin, Heidelberg: Springer.
- ▶ 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. (2022). "Improving the utility of Poisson-distributed, differentially private synthetic data via prior predictive truncation with an application to CDC WONDER." *Journal of Survey Statistics and Methodology*, **10**, 596–617.