Differential Privacy and Directional Noise Applied to Large Synthetic Medical Patient Data

Ryan McKay

Wednesday 27th February, 2019

0.1 What is Differential Privacy?

Differential Privacy (DP) is a mathematically rigorous framework for quantifying disclosure risk to an individual given their presence in a privatised data set. DP provides a guaranteed upper bound on information gain obtainable by an adversary with no assumptions about extra sources of information or computational power. [1]

The risk tolerance framing of DP can be controlled through its key parameters ϵ,δ and anecdotally understood by the following statements [1]:

- With high probability (1 - δ) DP can guarantee that an adversary could not gain more than e^{ϵ} times what she could have gained in expectation had any single observation been modified, added or removed.
- δ can be considered the probability of an uncontrolled privacy breach, this is insignificant near the mode and has more influence at the tails where probabilities may be near-0.
- ϵ is the worse case privacy loss when no uncontrolled privacy breach occurs and is more significant at higher concentrations of probability mass. [1]

0.2 What are the Ingredients of a Differential Privacy Scheme?

- Query - An arbitrary function of a data set eg $f(X) = X$ (Identity), $f(X) = \frac{1}{n}\sum_i X_i$ (Mean)
- A Random Mechanism - A function that takes in a query (data) and adds noise (typically symmetric) by some method.
- Parameters - We've discussed ϵ and δ already. Sensitivity is the maximum (global) change in query output by manipulating, adding or removing an observation from the data. This is a measure of the maximal change 'adjacent' queries. This value depends on the space of possible datasets, cannot be analytically calculated (local sensitivity, which does not guarantee DP).
- Properties - DP has two key properties that make it a flexible and useful framework.
  - Post-Processing guarantees that any output of a function of a DP dataset is also DP.
  - Composition guarantees that composed (ϵ_i,δ_i)-DP datasets will be ($\sum_i \epsilon_i, \sum_i \delta_i$)-DP. Consider the important case of sets of features composed from multiple datasets.
- Noise / Utility trade off - Generally adding more noise will increase the privacy of a dataset, but reduce its utility, or learnability. Much of DP research is dedicated to mechanisms and methods to reduce noise, while maintaining DP's rigorous privacy guarantees, or relaxing those guarantees strategically.

0.3 Solution

The canonical Gaussian (ϵ,δ) DP implementation is the addition of *iid* $N(0,O(\frac{log(\delta)}{\epsilon}*\text{Sensitivity}))$ noise to each element a dataset, which maintains DP by using the composition property. Notice noise added is symmetric, increases with sensitivity or inversely with ϵ, δ. Smaller parameters, increase noise and guaranteed privacy of the mechanism. A 2018 paper [3] describes a generalisation

[1] Included in my submission is a survey paper (unpublished) titled 'Differential Privacy for Private Data Release and Computation' that provides an approachable conceptual summary of differential Privacy frameworks

of this method using Matrix-variate Gaussian (MVG) Noise, that is adding a single matrix sample of noise from a Matrix-variate distribution [2]

Chanyaswad et al [3] proved that achieving (ϵ,δ) DP is conditioned only on a function of the sum of total variance of the mechanism which can be spread in any direction[3]. The best strategy is to optimise 'directional noise' to the direction and magnitude that will satisfy the DP constraint and maximise utility simultaneously.

Our solution is a prototype implementation of the MVG mechanism, one of the latest exciting results in DP for continuous data[4]. We have applied the above described DP schemes to preserve privacy of individual records in a synthetic, generated medical dataset [2]. We compare the empirical error of the output of 'directional noise' methods using the raw data and *iid* Gaussian mechanism as a baseline trained with machine learning algorithms Support Vector Regression, Principle Component Analysis (PCA).

Directional noise methods are presented in two general forms binary allocation strategy and derived directional noise. Binary allocation gives the majority of the precision budget (lowest noise) in the most useful directions (features) in equal amount and the remainder to the other directions[5]. Derived directional noise uses some function of derived directions and magnitudes of noise that aims to maximise some measure of utility. See appendix B for an implementation algorithm for the MVG mechanism with 'unimodal directional noise'[6].

0.4 Computational Complexity

Note: m, n are the dimensions of the matrix and s is the efficiency of the Gaussian sampler used.

- Symmetric Matrix $i.i.d$ Gaussian Mechanism has $\mathcal{O}(s\frac{n^2+n}{2})$.
- Matrix $i.i.d$ Gaussian Mechanism has $\mathcal{O}(smn)$.
- MVG Mechanism has $\mathcal{O}(m^3)$ [3] with m^3 factor for matrix multiplication, affine transformation sampling method and if using non derived directional methods such as eigenvalue decomposition (PCA) are also m^3.

0.5 Experimental Results

0.5.1 Private Sample Covariance / Symmetric matrix methods

Comparison: Identity query ($q(X) = X$) vs Centered Covariance Query ($p(X) = \frac{1}{n}X^TX$) For this experiment we compared the principal component residual sum of squares of the two methods, that is comparing the difference between eigenvalues of the raw and DP covariance matrices. Identity query method adds DP preserving noise to the raw data, then takes the sample covariance of this matrix. Centered Covariance query takes the sample covariance of the raw data and adds symmetric DP preserving noise. The metric remains stable for the identity query and decreases as the number of observations increases for the centered covariance query (see Apendix C).

[2]Matrix-variate distributions are generalisations of a multivariate distribution. They are specified by two covariance matrices, one for each rows and columns of a matrix.

[3]Though since this function is an inverse sum of squares See appendix A, the optimal noise added should be regularised and no direction can approach 0.

[4]Techniques in differential privacy tend to be solutions for continuous or categorical data. Using the composition property of DP we can easily compose different DP mechanisms to combine these

[5]Typically this uses prior knowledge about feature importance.

[6]Unimodal refers to the derivation of one direction of variance and setting the other to be iid identity covariance. This is typically when we wish to optimise directional noise for features, but not observations

0.5.2 Support Vector Regression (SVR)

We measure the utility retained by each Differentially Private Mechanism by taking 25 random samples from the mechanism and fitting a SVR 'RBF' kernel model to each with 'framingham score'[7] as the response and various patient observational measurements as the predictors. For each set of experiments we sample n independent observations from our larger dataset of 360,000 observations and set aside 15% each for parameter tuning and as a test holdout set to ensure our results are not due to over-fitting to the training set. The time complexity of each class of DP mechanism is reported in Appendix D and mean absolute error results can be found in Appendices E and F.

0.6 Conclusions

Overall ability of the DP Mechanisms to retain utility was encouraging, though they were open to notable variability in results. The Matrix-variate Gaussian (MVG) Mechanisms using binary allocation strategy performed the most consistently, while the derived directional variance MVG Mechanisms had the lowest error results, but the largest variability. The baseline model improved slightly as the sample size increased, but the DP Mechanism models did not demonstrate a clear improvement. It must be considered that for increases in the number of observations and features, the total noise added increases, however we can target that noise to less useful features. This relation may add to volatility, though sample size appears to balance its impact. Finally for modern standard we have used a relatively trivial example where more data may have severely diminished returns for utility and this raises the careful consideration of selecting the necessary sample size as to minimise noise added. Queries that include a $\frac{1}{n}$ term only benefit from an increase in sample size.

Further development of task targeted methods for directing and reducing noise will be require to generalised this procedure for broader machine learning, however given certain conditions these frameworks provide a flexible and robust framework for mathematically guaranteeing the privacy of individuals and may enable previously siloed data to be analysed and utilised to the benefit of all participating.

[7] A measure of the risk of cardiovascular disease in the next 10 years

Appendices

Appendix A

$\mathbf{X}$	database/dataset whose columns are data records and rows are attributes/features.
$\mathcal{MVG}_{m,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{\Psi})$	$m \times n$ matrix-variate Gaussian distribution with zero mean, the row-wise covariance $\mathbf{\Sigma}$, and the column-wise covariance $\mathbf{\Psi}$.
$f(\mathbf{X}) \in \mathbb{R}^{m \times n}$	matrix-valued query function
r	$\min\{m, n\}$
H_r	generalized harmonic numbers of order r
$H_{r,1/2}$	generalized harmonic numbers of order r of $1/2$
γ	$\sup_{\mathbf{X}} \|f(\mathbf{X})\|_F$
$\zeta(\delta)$	$2\sqrt{-mn \ln \delta - 2 \ln \delta + mn}$
$\boldsymbol{\sigma}(\mathbf{\Sigma}^{-1})$	vector of non-increasing singular values of $\mathbf{\Sigma}^{-1}$
$\boldsymbol{\sigma}(\mathbf{\Psi}^{-1})$	vector of non-increasing singular values of $\mathbf{\Psi}^{-1}$

Table 1: Notations for the differential privacy analysis.

Theorem 2. *Let* $\boldsymbol{\sigma}(\mathbf{\Sigma}^{-1}) = [\sigma_1(\mathbf{\Sigma}^{-1}), \ldots, \sigma_m(\mathbf{\Sigma}^{-1})]^T$ *and* $\boldsymbol{\sigma}(\mathbf{\Psi}^{-1}) = [\sigma_1(\mathbf{\Psi}^{-1}), \ldots, \sigma_n(\mathbf{\Psi}^{-1})]^T$ *be the vectors of non-increasingly ordered singular values of* $\mathbf{\Sigma}^{-1}$ *and* $\mathbf{\Psi}^{-1}$*, respectively, and let the relevant variables be defined according to Table 1. Then, the MVG mechanism guarantees* (ϵ, δ)*-differential privacy if* $\mathbf{\Sigma}$ *and* $\mathbf{\Psi}$ *satisfy the following condition,*

$$\left\|\boldsymbol{\sigma}(\mathbf{\Sigma}^{-1})\right\|_2 \left\|\boldsymbol{\sigma}(\mathbf{\Psi}^{-1})\right\|_2 \leq \frac{(-\beta + \sqrt{\beta^2 + 8\alpha\epsilon})^2}{4\alpha^2}, \tag{1}$$

where $\alpha = [H_r + H_{r,1/2}]\gamma^2 + 2H_r\gamma s_2(f)$*, and* $\beta = 2(mn)^{1/4}\zeta(\delta)H_r s_2(f)$*.*

Appendix B

Algorithm 1 MVG mechanism with unimodal directional noise.

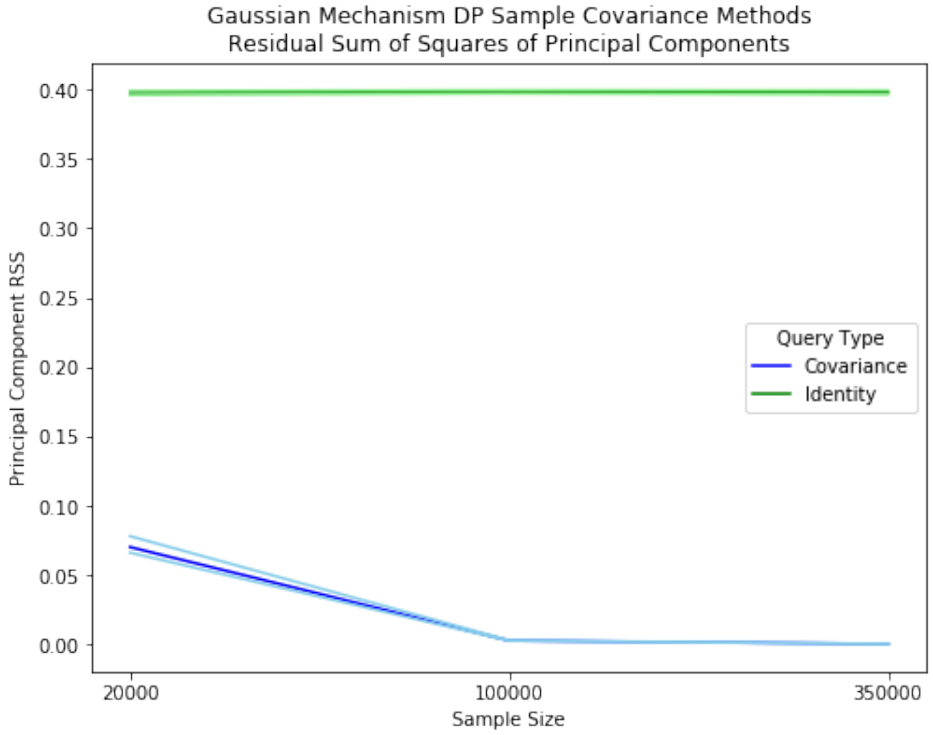
Input: (a) privacy parameters: ϵ, δ; (b) the query function and its sensitivity: $f(\mathbf{X}) \in \mathbb{R}^{m \times n}, s_2(f)$; (c) the precision allocation strategy $\boldsymbol{\theta} \in (0,1)^m : |\boldsymbol{\theta}|_1 = 1$; and (d) the m directions of the row-wise noise $\mathbf{W}_{\mathbf{\Sigma}} \in \mathbb{R}^{m \times m}$.

1. Compute α and β (cf. Theorem 2).
2. Compute the precision budget $P = \frac{(-\beta + \sqrt{\beta^2 + 8\alpha\epsilon})^4}{16\alpha^4 n}$.
3. **for** $i = 1, \ldots, m$:
   (a) Set $p_i = \theta_i P$.
   (b) Compute the i^{th} direction's variance, $\sigma_i(\mathbf{\Sigma}) = 1/\sqrt{p_i}$.
4. Form the diagonal matrix $\mathbf{\Lambda}_{\mathbf{\Sigma}} = diag([\sigma_1(\mathbf{\Sigma}), \ldots, \sigma_m(\mathbf{\Sigma})])$.
5. Derive the covariance matrix: $\mathbf{\Sigma} = \mathbf{W}_{\mathbf{\Sigma}}\mathbf{\Lambda}_{\mathbf{\Sigma}}\mathbf{W}_{\mathbf{\Sigma}}^T$.
6. Draw a matrix-valued noise $\mathbf{Z}$ from $\mathcal{MVG}_{m,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I})$.

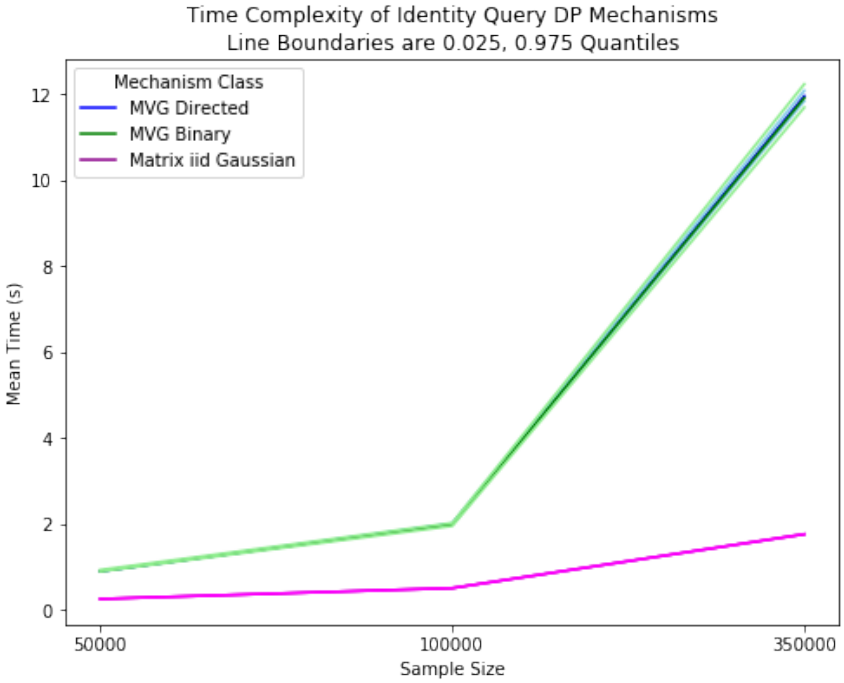
Output: $f(\mathbf{X}) + \mathbf{Z}$.

Appendix C

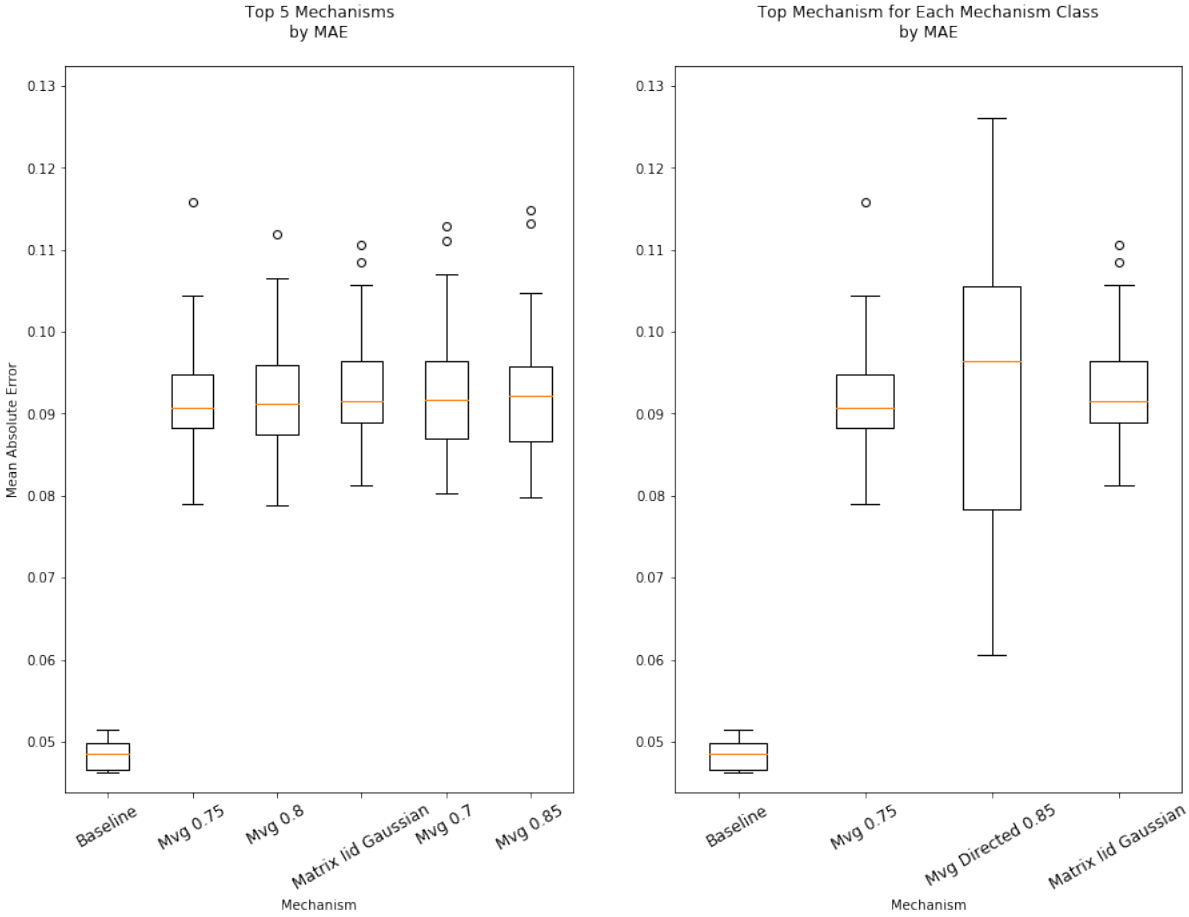
	sample size	query	metric	result	0.025 quantile	0.975 quantile	mechanism runtime (s)
0	20000	covariance	principal component RSS	0.070118	0.066097	0.077968	0.000474
1	20000	identity	principal component RSS	0.397775	0.396027	0.399180	0.007805
2	100000	covariance	principal component RSS	0.002646	0.002525	0.002841	0.000484
3	100000	identity	principal component RSS	0.398695	0.397052	0.399731	0.007762
4	350000	covariance	principal component RSS	0.000286	0.000260	0.000299	0.000482
5	350000	identity	principal component RSS	0.398079	0.396392	0.399569	0.007831



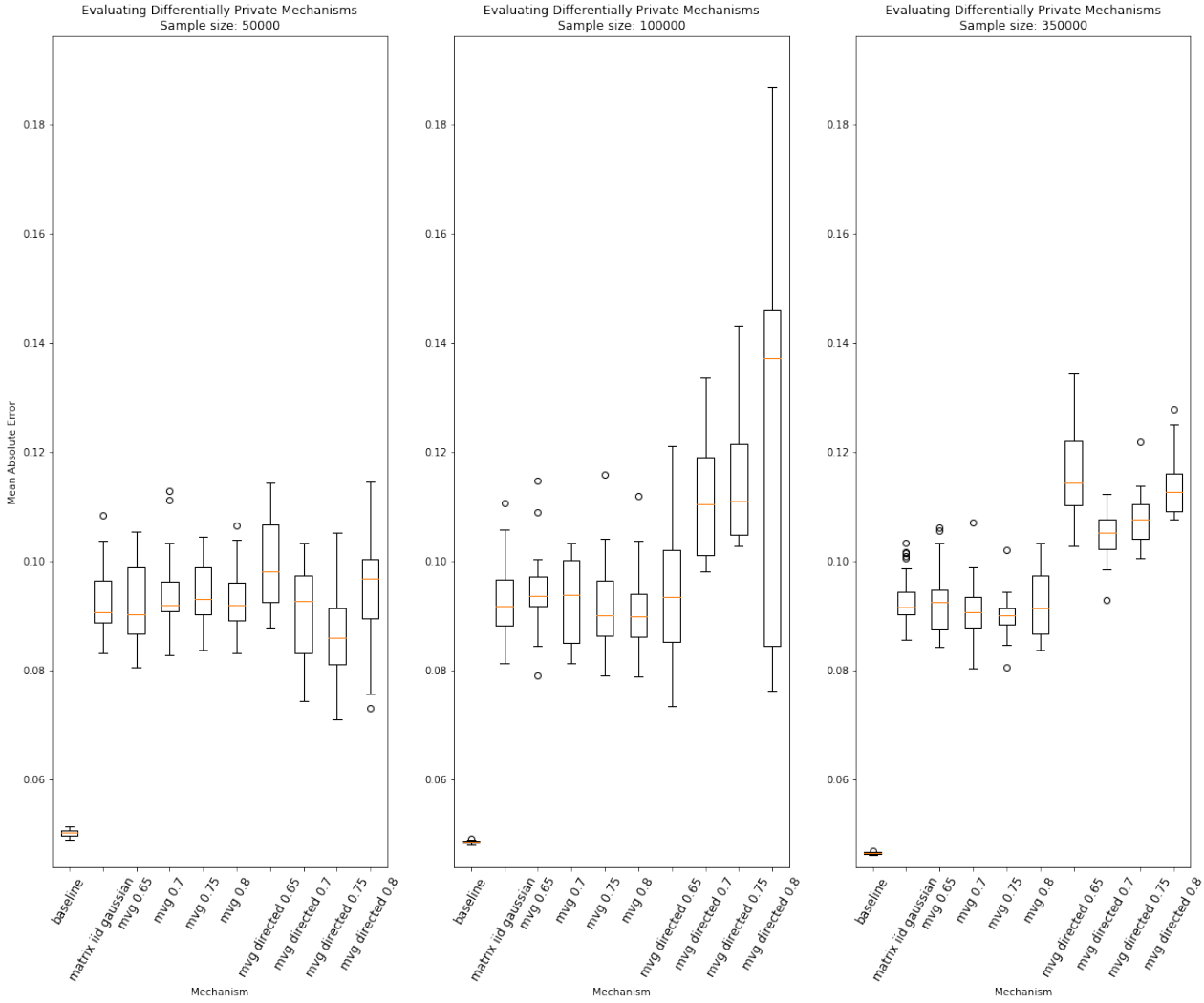
Appendix D



Appendix E



Appendix F


Evaluating Differentially Private Mechanisms
Sample size: 50000
Evaluating Differentially Private Mechanisms
Sample size: 100000
Evaluating Differentially Private Mechanisms
Sample size: 350000
Mean Absolute Error
0.06
0.08
0.10
0.12
0.14
0.16
0.18
baseline
matrix iid gaussian
mvg 0.65
mvg 0.7
mvg 0.75
mvg 0.8
mvg directed 0.65
mvg directed 0.7
mvg directed 0.75
mvg directed 0.8
Mechanism

References

[1] Le Nguyen Hoang. Interpretation of the ϵ and δs of differential privacy, 2017.

[2] Andre Quina, Carlton Duffett, Chris Moesel, Dylan Hall, Joseph Nichols, Mark Kramer, Jason Walonoski, Kudakwashe Dube, Scott McLachlan, and Thomas Gallagher. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, 08 2017.

[3] Joshua Snoke and Aleksandra Slavkovic. pmse mechanism: Differentially private synthetic data with maximal distributional similarity. 05 2018.