# Differential Privacy Applied on Menstruation Data

**Nathan Dennis, Raima Islam, Lia Zheng, Shibani Rana**

May 8, 2025

**Abstract**

This paper explores the application of differential privacy methods to the release and analysis of menstruation data from period tracking apps, focusing on ensuring individual privacy while still providing useful insights. Through using noisy means, histograms, and differentially private deep learning models through DP-SGD (Adam variant) we simulate how period tracking apps could release privacy preserving data such as average cycle length and menses lengths. We apply DP techniques like Laplace noise and DP-SGD to generate private aggregates and predictive models, assessing the privacy accuracy trade off through these approaches. The results highlight key trade offs between privacy and accuracy, where as expected stronger privacy with smaller $\epsilon$ values results in more noisy results and worsened model performance. More moderate privacy settings like $\epsilon \geq 1$ can offer a more balanced privacy accuracy tradeoff, but the challenge on choosing $\epsilon$ still remains. This work contributes to the ongoing challenge of balancing privacy and accuracy in the release of health related data, offering more findings into how DP can be applied to sensitive menstrual health data.

## 1 Introduction

Period tracking apps introduce a unique challenge in the realm of differentially private data releases, ensuring that the analysis of menstruation data remains both accurate and useful for users and third parties. In order for data to be useful to third parties, it is essential to be able to accurately predict attributes like length of cycle, while for users the ability to compare their own menstruation patterns with those of the entire population is equally important. This paper explores methods in differential privacy that allow releases of such data while maintaining individual level privacy.

Through the use of noisy means, noisy histograms, and deep learning models with differentially private stochastic gradient descent (DP-SGD), we simulate how a period tracking app could share menstruation data to users and third parties. The ability to release information and compute predictions about menstruation data (such as average cycle length, menses length, and bleeding intensity) in a differentially private way is an invaluable opportunity for individuals to learn about their health while preventing the misuse of such sensitive data.

## 2 Background and Related Work

Differential privacy is a technique that is designed to provide privacy guarantees for individuals within a dataset. The core idea of DP is to introduce noise in such a way that individual data points can't be distinguished, or at least with high confidence, even when combining the data with other external datasets. The level of noise added is controlled by a key parameter, $\epsilon$, which highlights the trade off between data accuracy and privacy protection.

In general, DP has found widespread applications in the release of aggregate data, such as census data and public health statistics. For example, the U.S. Census Bureau used differential privacy techniques in the 2020 census to protect individual responses while still enabling valuable insights for public use [5]. By applying DP, the Census Bureau was able to balance privacy concerns with the need for accurate and reliable demographic data, which was an important advancement for these types of government surveys.

The healthcare domain has also begun using differential privacy, where recent work has experimented with applying DP to health related data such as public health records. Studies have shown that DP can be used to aggregate statistics from sensitive data with strong privacy guarantees, while maintaining useful accuracy [3]. This highlights the potential usability of differential privacy in the medical field, using basic differential privacy to privatize aggregate statistics releases. This also includes applications with sensitive medical data, such as genomic and wearable device data, where DP has been used to preserve privacy while enabling meaningful analysis [10].

More recent work has also explored the challenges of applying differential to machine learning tasks. For instance, Jayarnaman and Evans evaluated DP machine learning in practice and found that achieving meaningful privacy guarantees often requires sacrificing model utility, and that current DP training mechanisms struggle to provide acceptable privacy utility tradeoffs, particularly in complex learning tasks.[8]. This result highlights the difficulty of achieving strong privacy guarantees without significant reductions in model accuracy. Building on this, recent work has also explored applying differential private machine learning in healthcare specific contexts, specifically training deep learning models on medical datasets using differential privacy. For example, Dopamine is a system that applies federated learning with differentially private stochastic gradient descent (DP-SGD) to medical data. By combining secure aggregation with DP training, the systems achieves better accuracy and privacy trade offs compared to other DP approaches. It was tested on a diabetic retinopathy classification task and achieved competitive performance under string privacy constraints, showing DP's potential in machine learning [11].

Despite the advancements in DP for medical datasets, there has been more limited exploration of how differential privacy can be effectively applied to more specific forms of sensitive medical data, such as menstruation data. This paper aims to address this problem, focusing specifically on using DP methods on menstruation data in a way that balances privacy with accuracy for both individual users and third party researchers.

# 3 Data

## 3.1 Dataset Description

We used a dataset titled "Menstrual Cycle Data" which originated from a 2012 randomized clinical trial conducted by Richard J. Fehring at Marquette University [4]. The study originally wanted to compare the efficacy and acceptability of two internet support natural family planning (NFP) methods. The dataset is comprised of 1650 rows, each detailing a single menstrual cycle and its related information including variables such as cycle length, luteal phase length, menses length, bleeding intensity and fertility indicators, as well as participant demographics like ethnicity, age, and BMI.

## 3.2 Data Cleaning

The dataset consisted of entries that corresponded to a single menstruation cycle, with multiple entries per individual. To clean the data, we split the dataset into two datasets—one for values per cycle (1650 entries) and one for values per person (159 entries). The values per person were averages of all of their cycles as well as their demographic information. For instance, if one person had 20 entries in the cycles dataset, their entry in the person dataset would contain average cycle length, average menses length, and average bleeding intensity of all 20 of their cycles.

Many attributes were missing for different individuals and cycles in the original dataset. We attempted imputation methods such as mean imputation, however the results did not accurately reflect real populations since the dataset was relatively small to begin with. We instead decided to drop missing rows since the features that we wanted to produce statistics for had little to no missing rows.

# 4 Methodology

In this project, we explore two complementary approaches to study sensitive menstruation cycle data, while hoping to preserve user privacy. First, by generating differentially private aggregates, we can release useful statistics such as average cycle length without exposing individual records for users to compare themselves to the population averages. We can also release differentially private distributions for users to understand where they lie on the population distribution. Second, by applying DP-SGD to deep learning models we hope to learn how to make predictions while ensuring no single users data has significantly influenced the model. Together, these methods can allow us to balance the need for meaningful health insights but also with the goal of protecting individual privacy.

## 4.1 Part I: Private Aggregates & Histograms

This first approach reflects much of what we have done in class. The objective with this approach is to extract meaningful insights from menstruation app data—e.g. average cycle lengths, the distribution of these averages across users, stratified into various demographic factors—while ensuring individual user level privacy. We are able to compute DP means across the user base, as well as produce DP-histograms.

We consider DP for adjacent datasets, where adjacency is defined as the absence or addition of one user. For all statistics (Cycle Length, Menses Length, Bleeding Intensity, and Day of Ovulation), we used Laplace noise based on the global sensitivity. Cycle and Menses Length are calculated by number of days and Bleeding Intensity is calculated by summing the scores for each day of Menses (3 being the highest intensity). For the relevant variables, we computed the overall DP means and DP means stratified by demographics (Ethnicity, BMI, Age, and Birth Control Method). We also computed DP distributions of these variables, along with the distributions of their means (also stratified by demographics) and variances. We used an $\epsilon$ range of [0.1, 0.5, 1, 5]. Different $\epsilon$ values provided different levels of interpretable results, as well as levels of accuracy of the histogram. We checked the accuracy by comparing the DP mean values with the true mean values and DP histograms with the true histograms.

Note that we did not do this via synthetic data generation—instead, we generate DP-counts for each histogram bin. We acknowledge that this method does result in some bias, since there cannot be negative counts in histogram bins. We use a bin-first method, where we first determine the number of bins to use in our visualization, then add noise. The reason for this is that most of the statistics will fall within a certain range (i.e. period cycles will likely last between 20-35 days), and using days as a unit, there are naturally some binning divisions that are more intuitive than others (i.e. 10 days would be too large for 1 bin).

Since we use two datasets—`cycle_data_by_ID.csv` and `cycle_data_filled.csv`, the sensitivities are different across the two datasets, since in `cycle_data_by_ID.csv`, each user contributes at most 1 entry. In `cycle_data_filled.csv`, one user contributes over 30 entries, meaning that the sensitivity is much greater. As a result, the accuracy for the histograms using this dataset are much lower.

### 4.2 Part II: Predictive Modeling with DP-SGD

#### 4.2.1 Differential Privacy Implementation

We applied differential privacy to neural networks using the Opacus library, which implements DP-SGD, using the Adam variant. In this library, per-sample gradients are clipped to some fixed norm which limits individual gradient influence, then Gaussian noise is added to ensure privacy in each batch. We specifically used the DP-SGD Adam variant which was available in the library, which builds on Adam optimization (including RMSProp with adaptive learning rates and Momentum) while preserving the privacy guarantees of DP-SGD. It applies the same principles of DP-SGD (clipping and noise addition) to Adam's gradient update steps.

When using the specific `make_private_with_epsilon` method that is from Opacus and `PrivacyEngine`, the library automatically determines the appropriate noise multiplier to achieve the target total privacy budget that we can manually decide, $\epsilon$ and $\delta$, over some fixed number of training epochs. The total privacy cost is not evenly split per epoch, instead each batch contributes a small amount to the overall privacy loss as the training repeatedly sees the data for each epoch. Opacus uses Renyi Differential Privacy (RDP) accounting to track this cumulative loss during training, and the final $\epsilon$ and $\delta$ is computed after all the training steps. The target $\epsilon$ was varied for analysis, but the target $\delta$ was always set at $1e-5$.

When comparing the privacy-accuracy trade-off, we used the following $\epsilon$ values: 0.1, 0.3, 0.5, 1, 5, 10. We chose these values since smaller $\epsilon$ values would promote stronger privacy by adding more noise to the gradients, but potentially at the cost of decreased accuracy. However, higher $\epsilon$, such as 5 or 10, inject less noise, potentially preserving more model performance at the cost of weaker privacy guarantees. Furthermore, smaller $\epsilon$ values could encourage more varied accuracies and less consistency with the greater noise added, which we will also analyze.

#### 4.2.2 Model Approaches and Architectures

We developed two different models with DP-SGD. For the first model, we applied a regression mechanism to predict continuous values of *Length of Cycle*, which is a biologically relevant indicator of how severe or minor it is, which can be predictive of factors such as PCOS, fertility, or hormonal problems. This feature might serve as an informative measure for any sort of period-tracking or digital health applications. For this task, we built a three-layer feedforward neural network with ReLU activation function, which includes hidden layers with 64, 32, 16 units, followed by dropout regularization to fix common problems in neural networks: vanishing and exploding gradients, overfitting. The input features for this problem setup included both numerical and categorical columns, such as age, BMI, bleeding intensity, ethnicity, etc. which were processed with one-hot encoding. To ensure DP, we have utilized the DP-Adam optimizer, which performs

per-sample gradient clipping and Gaussian noise addition, adhering to the theoretical rules of DP-SGD. This model was trained on a variety of privacy budget values, and performance was evaluated by finding their corresponding Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values.
For the second model, we trained models to predict `Medvits`, which indicates whether an individual is currently taking medications or vitamins (1 = yes, 0 = no). This task was chosen due to its relevance in personal health behavior and potential value to third party groups such as advertisers and researchers. Advertisers for example would be interested in predicting whether or not an individual buys supplements or medications since they could send them targeted ads for these health-related products. We used a 2 layer feed forward neural network with ReLU activations, where the first hidden layer had 64 units and second had 32 units. The final layer consisted of a single output node with a sigmoid activation for binary classification. To handle class imbalance in `Medvits`, which has nearly a 3:1 ratio in favor of the 'yes' class (takes medications or vitamins), we used a class-weighted binary cross entropy loss function. The model was trained using DP-Adam from the Opacus library, which uses stochastic gradient descent with Gaussian noise addition and per-sample gradient clipping as explained above.

#### 4.2.3 Feature Selection

For the regression model, our objective was to detect the Length of Cycle, we selected features comprising both physiological and demographic-level information, along with menstrual health features. We have selected Age, which is a determining factor in how heavy or light your flow will be, along with understanding cycle duration patterns [12]. For BMI, it is correlated to the menstrual cycle length, variability, depending on how low or high the number is, and reproductive ability, to understand regularities and irregularities [6]. Ethnicity, as well, plays a crucial role, where certain period management practices and beliefs affect our cycles [6]. For the other categories, such as mean bleeding intensity, length of menses, estimated day of ovulation, number of pregnancies, reproductive category, etc, are important for understanding underlying health conditions [7].

For the second binary classification model, we selected eight numeric features that capture core aspects of menstrual cycle and individual health profiles, all of which are potentially important for predicting medication or vitamin use. First, we used cycle length including length of cycle for every individual and their length of luteal phase since variations in menstrual cycle or luteal phase has been linked to hormonal imbalances and potential reproductive health issues, which may prompt medical treatment or supplementation [1]. Similarly we included length of menses, which refers to the duration of the menstrual bleeding period. Longer menstrual bleeding can lead to increased blood loss, which could contribute to iron deficiency [13]. Iron deficiency is commonly addressed with vitamin or iron supplements, which individuals may take to prevent or treat iron deficiency and improve overall health. We also included variables highlighting mean menstrual cycle length and mean menstrual bleeding period for each individual. These variables can provide deeper insights into reproductive health and may influence the likelihood of vitamin usage, particularly for individuals experiencing irregular cycles or prolonged bleeding, which are often linked to conditions requiring supplementation.

We also included the estimated day of ovulation since ovulation is associated with a range of physical and mood changes. Individuals experiences ovulation related symptoms, such as bloating and bleeding, may be more likely to take vitamins to support their reproductive health and cycle regulation [2]. This information could help advertisers or researchers better predict vitamin or medication intake behaviors and identify periods of increased health related product interest.

As for demographic factors, we included age first in the model because older individuals are more likely to take medications or vitamins than younger people. They are also more likely to engage in "polypharmacy" which is defined as the use of five or more medications at the same time, due to multiple chronic health conditions which often leads to increased use of supplements and vitamins to manage or counteract the effects of medication [14]. BMI was included in the model since research has shown that multivitamin use correlates with BMI, with individuals in lower BMI categories being more likely to use multivitamins. In a study on multivitamin use it was found that multivitamin use decreased with increasing BMI, potentially due to lower perceived health needs for vitamin supplements in higher BMI groups [9].

# 5 Results

## 5.1 Part I: Private Aggregate Results

To show the results of our first method, we will use the example of Cycle Length. Results for other statistics using the same method can be found in the Appendix, in Section 7. Very small values of $\epsilon$ resulted in unrealistic statistics and absurd-looking graphs—in the context of menstruation cycles, the graphs did not make sense, specifically as menstruation cycles are biologically constrained within certain ranges, and there was too much noise added to preserve accuracy, as each random generation would produce wildly varying graphs. We found that an $\epsilon$ value of 1 was the most useful for all variables, while being a value that probabilistically protects user privacy.

### 5.1.1 DP Means

Table 1 shows the population cycle length mean for different values of epsilon along with the true mean. Table 10 shows the same means stratified by demographic groups. Note that the privacy budget is split evenly per group. The DP releases of population averages would allow users to compare their own averages to the "norm". For cycle length, users would be able to identify whether they deviate significantly from the population. They can also identify how their demographic group affects their cycle length. We can see the as the privacy budget increases, the patterns for for demographic groups match the true patterns. However, the demographic group imbalance in the dataset creates an issue where the statistics ordering in the dataset do not reflect real-world patterns. Also, since the privacy budget is split evenly for each group histogram, the amount of random noise added can disproportionately affect each group and change the final ordering.

| | True | $\epsilon = 0.1$ | $\epsilon = 0.5$ | $\epsilon = 1$ | $\epsilon = 5$ |
|---|---|---|---|---|---|
| Mean | 29.32 | 23.70 | 28.65 | 30.02 | 29.91 |

Table 1: Cycle Length Means

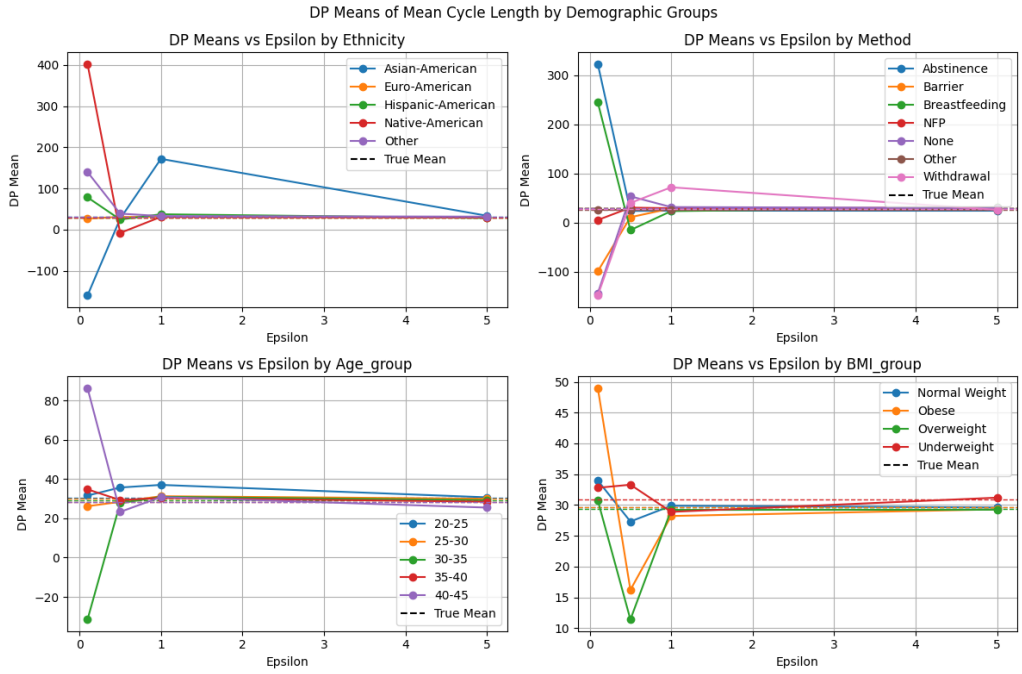



Figure 1: Cycle Length Means by Demographics

### 5.1.2 DP Histograms

These DP histograms could be communicated to app users, allowing them to see where they fall on the privatized distribution—note the density plots, not counts. The mean is directly taken from `cycle_data_by_ID.csv`. The standard deviations are calculated using `cycle_data_by_FILLED.csv`, where we find the standard deviation over every single one of a user's entry. One potential issue with this method is response bias, since

users that only have 1 entry have a standard deviation of 0. We generated the DP histograms for different privacy budgets to compare the privacy-accuracy tradeoffs.

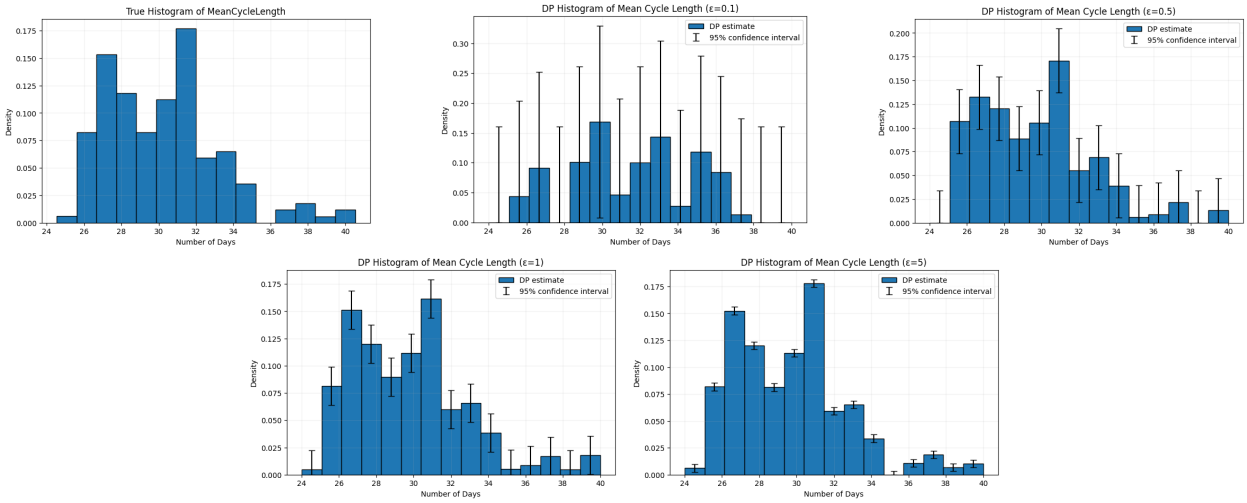



Figure 2: Mean Cycle Length Distributions

For the distribution of all cycle lengths, the data is centered around 30 days. The distribution of mean cycle lengths per person is also centered around 30 days. Users can compare their cycle lengths to the distribution to understand whether their own cycles are significantly outside the norm.

Note the 95% confidence intervals that show the uncertainty regarding the true sample statistics. These confidence intervals estimate what the true histograms look like. The wider the interval, the less certainty there is about the true value. We can see a clear privacy-accuracy trade-off; as epsilon increases, there is less privacy by definition of DP, but the error bars shrink significantly. For $\epsilon = 0.1$, and even $\epsilon = 0.5$, the error bars are so large that the DP histograms are not useful for providing insight. However, we can see that there are still useful results with $\epsilon = 1, 5$. Since an $\epsilon = 1$ means that the denial ratio is $e$, this is very reasonable performance. With an even larger database, as popular period apps like Flo are used by 70 million users monthly, these error bars will become negligible. Even in our dataset of 159 users, the error bars are already small enough for accuracy for $\epsilon \geq 1$.

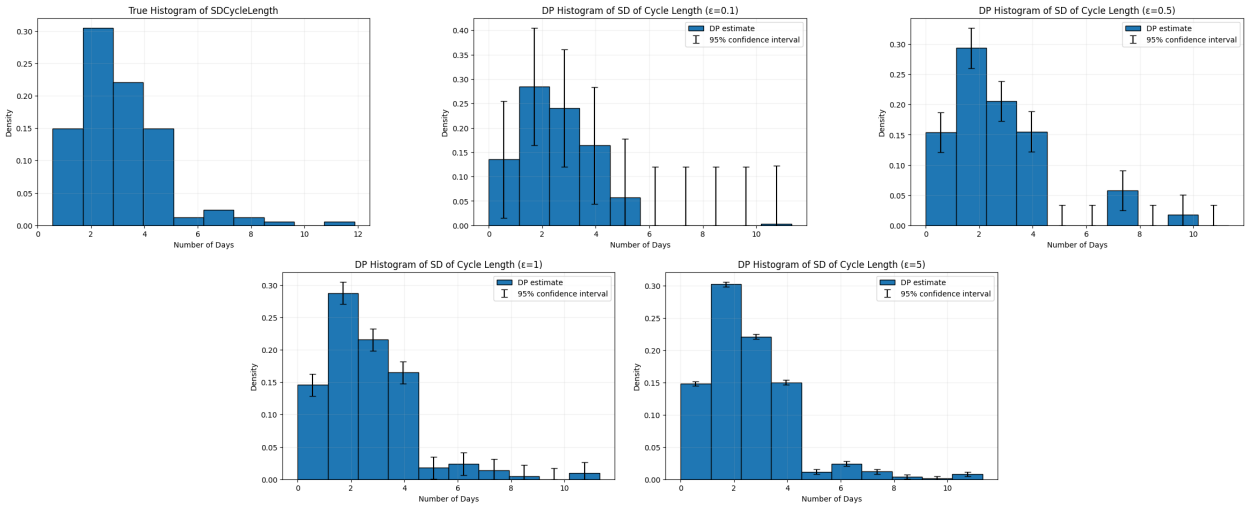



Figure 3: Cycle Length Variance Distributions

The distribution of standard deviations for cycle lengths allows user to compare how much their own cycle lengths change to the population. For instance, if you have a cycle length standard deviation of 10 days, you may be slightly concerned since you are on the tail of the distribution (which may be a helpful

sign to seek medical advice), since it seems most people have cycle standard deviations between 0 and 4 days. For standard deviation, the error bars are generally smaller than the mean cycle length for the same $\epsilon$, which aligns with expectations, since standard deviation of cycle length should have a smaller range than mean cycle length, thus a smaller global sensitivity.

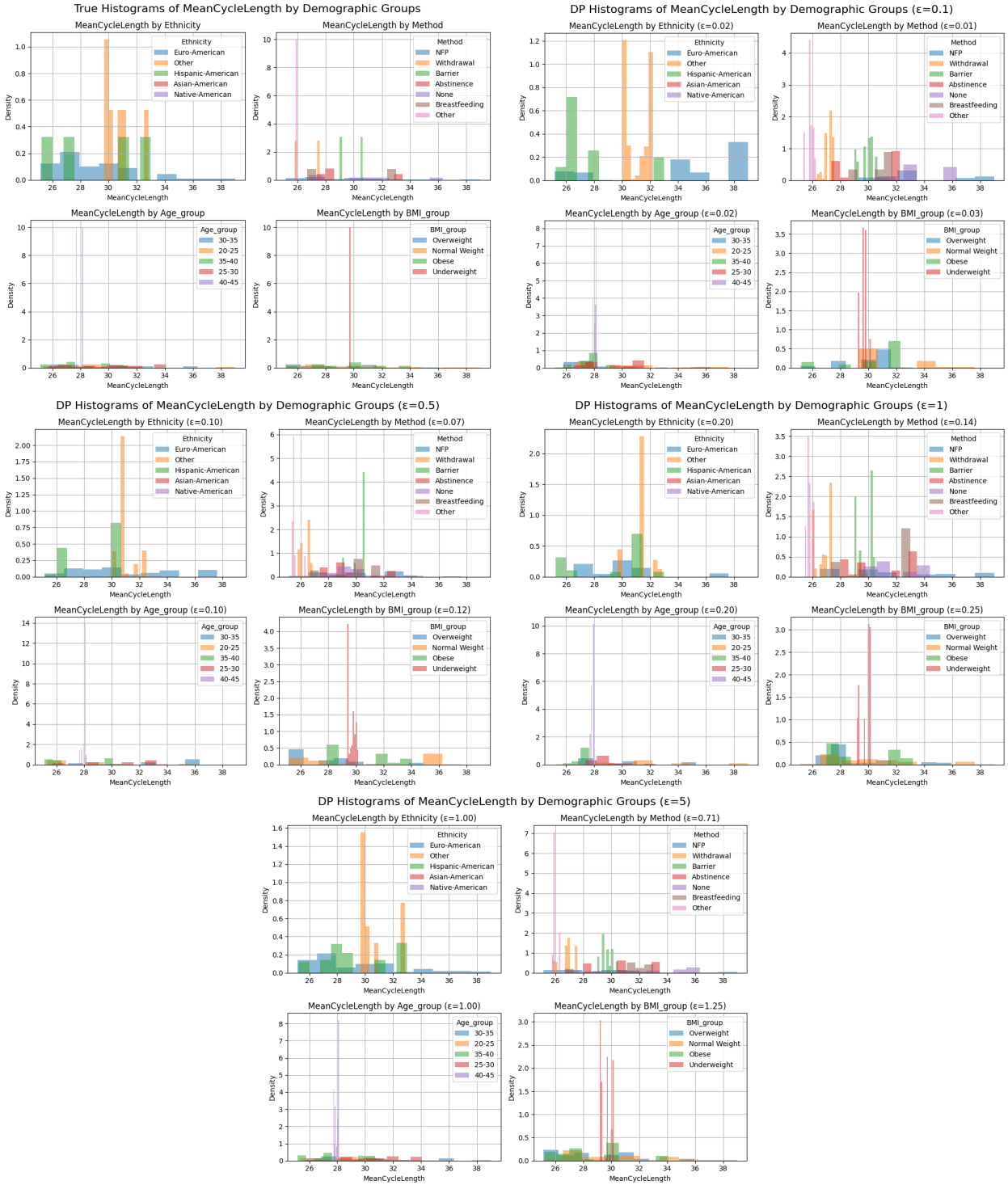



Figure 4: Cycle Length Distributions by Demographics

Finally, the mean cycle length distributions stratified by demographics allows users to identify where they lie on the distributions that match their own demographics. The privacy budget is split among the groups in each demographic attribute. Since different demographic attributes can affect menstruation, these distributions can allow users to narrow down the distribution for their own purposes, while still providing privacy. These distributions were much less evenly distributed due to the class imbalances in demographic groups. With more data, these histograms would be more accurate and useful. However, for proof of concept, we can see that different groups have different distributions for mean cycle length. Therefore, users should be able to see where they lie on distributions that match their own demographics. Significant deviations can motivate people to seek medical advice.

### 5.2 Part II: DP-SGD Model Results

In this next section we present the DP-SGD (Adam variant) model results. We first analyze the regression task where we developed a deep learning model to predict Length of Cycle (basically the length of a period) as described previously. We then present the binary classification model results predicting whether or not an individual takes vitamins. For the regression task we evaluate using MSE/RMSE and for the classification task we observe both accuracy and common metrics including TPR, FPR, FNR, and TNR. Furthermore since SGD is stochastic and the noise applied to gradients would be as well, we computed across 20 simulation runs rather than a single run to get more accurate aggregate results. We plot the average MSE/RMSE/accuracy across the 20 simulation runs and fill with the standard deviation across the runs for better analysis.

For the regression task, we experimented with differential privacy to understand its effects on the model's capability of predicting Length of Cycle. We have trained the complex neural network mentioned above on multiple privacy budgets, using DP-Adam optimizer. our evaluation focused on MSE and RMSE as evaluation metrics, computed across 30 epochs per $\epsilon$ value.

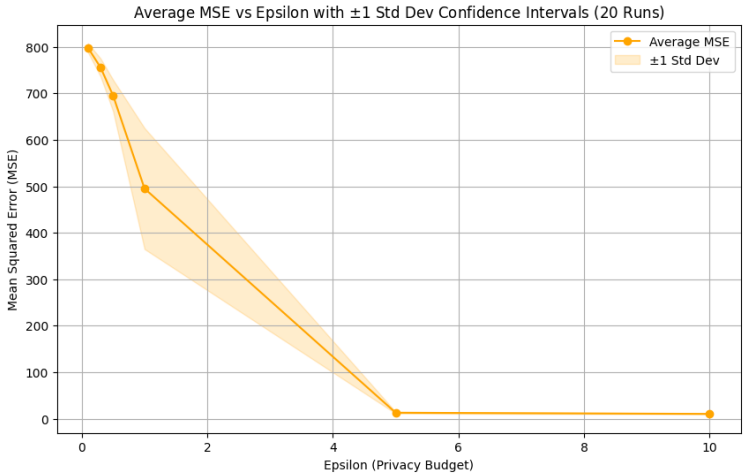



Figure 5: DP-SGD performance on $\epsilon$ values for Regression (MSE)

Figure 5 demonstrates the average test MSE scores for each $\epsilon$ value with a +/ standard deviation confidence intervals. As we have learned from class, we can see that the error increases as the epsilon value gets tighter, which is reflective of the privacy-accuracy tradeoff. For the $\epsilon$ values between 0.1-0.5, the MSE is exponentially high and very random, stipulating that the addition of noise is strongly correlated to the model's stability. Another interesting observation was that for higher $\epsilon$ values, for e,g. 5 or more onwards, the MSE converges to a much lower value, almost approaching baseline results.

For RMSE, we can see that from Figure 6, it almost follows the same trend where at $\epsilon$=10, the RMSE is very low (3.2), but for $\epsilon$=0.1, the RMSE is very high (exceeding 27). So, as we decrease the privacy guarantees, the performance improves vastly. For the confidence interval, this region shows a large variance for the smaller values, highlighting the instability of strong privacy guarantees. This mirrors the above plot but with a smaller y-axis limit since we use root mean squared error compared to regular MSE.

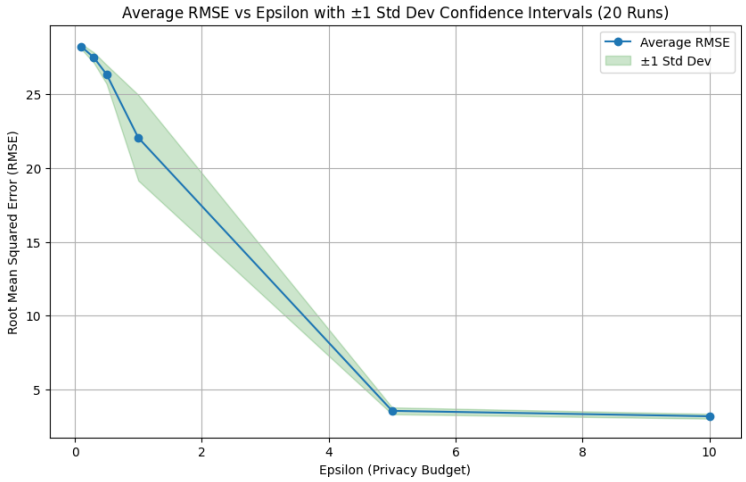



Figure 6: DP-SGD performance on $\epsilon$ values for Regression (RMSE)

Our results initially gave us some good direction on how to design the analysis. We experimented with in both DP vs. non-DP mode with our neural network model. Non-DP achieved an RMSE of 2.21 and 4.052 for DP-enabled model. But as we can see, with higher $\epsilon$ values, we can significantly lower the RMSE values, getting benchmark performances, but we still need to ensure privacy protection to such sensitive data.

For the binary classification task, we applied differential privacy to our neural network model with DP-SGD, Adam variant, across a range of $\epsilon$ values to assess accuracy and other metrics. The first plot, 7, shows the average test accuracy of the model as a function of $\epsilon$, with shaded $\pm 1$ standard deviation bounds from the 20 simulations. For comparison, the dashed red line highlights the average test accuracy for the non-DP Adam model across all 20 simulation runs, which serves as a baseline for the accuracy without differential privacy.

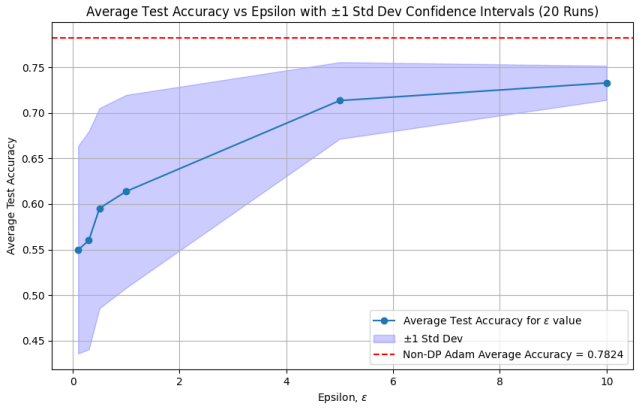



Figure 7: Accuracy Across Epsilons

As expected, we can see that the accuracy decreases as $\epsilon$ decreases, indicating the trade off between privacy and accuracy. The non-DP Adam model consistently outperforms the DP models, with a slight drop in average accuracy as $\epsilon$ values approach 5 and 10 which would be softer privacy guarantees compared to smaller $\epsilon$'s. For smaller $\epsilon$ values the performance gap widens, highlighting the impact of differential privacy on accuracy where greater noise results in far less accuracy compared to the baseline. Also, for smaller $\epsilon$ values the variation of accuracies depicted by the standard deviation is much larger for smaller values of $\epsilon$, indicating more unpredictable results with smaller values of $\epsilon$. The standard deviation for higher values of

$\epsilon$ are smaller and the results seem more predictable and consistent across runs.

In Figure 8, we observe the performance across $\epsilon$ values measured in terms of True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and True Negative Rate (TNR), compared to the baseline DP-Adam model. These metrics can provide us with a deeper understanding of the models performance when applying differential privacy. It would be important to visualize these metrics to observe differences in the way the model classifies both positive and negative cases (taking medications/vitamins or not).

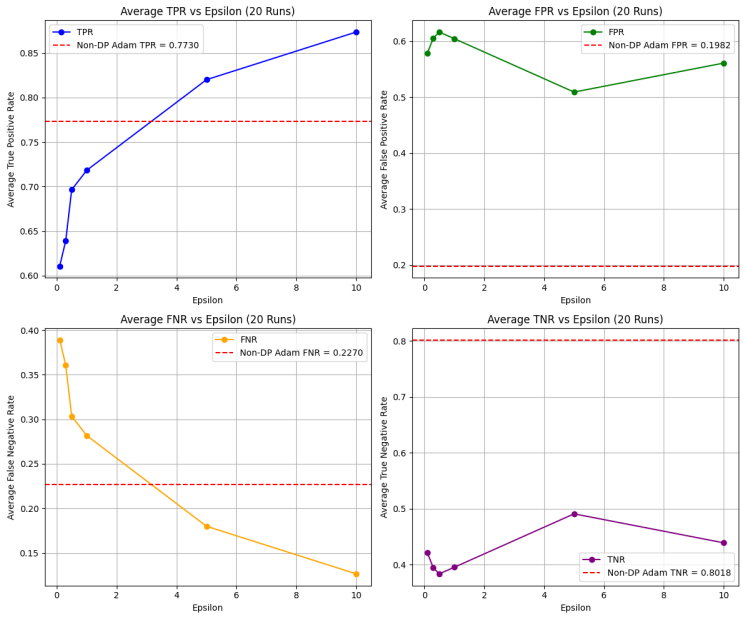



Figure 8: Metrics Across Epsilons

The top left plot highlights the TPR across $\epsilon$, which measures the proportion of actual positives correctly identified (individuals who do take medications/vitamins). As $\epsilon$ decreases, privacy increases, and the TPR gradually declines. A higher TPR indicates better identification of positive cases, but the model's ability to identify these positive cases is worsened with stronger privacy guarantees and lower $\epsilon$ values. We can see the opposite happen in the FNR case, which is the complement of the TPR which measures how often positive cases are missed, which increases as $\epsilon$ decreases. This again highlights the trade off between privacy and accuracy, stronger privacy (lower $\epsilon$) leads to more missed detections of individuals who take medications or vitamins. Compared to the baseline DP-Adam model, it lies near the middle of the privacy curve where higher $\epsilon$ values actually achieved a greater TPR and smaller FNR, but smaller $\epsilon$ had a smaller TPR and higher FNR than the baseline.

Next, the top right plot shows the FPR, which captures how often the model incorrectly classifies a negative case as positive (predicting someone takes medication or vitamins when they don't), shows more irregular behavior. For example, $\epsilon = 5$ results in the smallest FPR and $\epsilon = 0.5$ had the highest FPR. This suggests that FPR may be more sensitive to randomness or training instability under DP-SGD and doesn't follow a simple or montonic trend. The same can be said for the TNR, the complement of FPR, which mirrors this. While smaller $\epsilon$ tend to lead to smaller TNR (worse identification of negatives), the pattern is less predictable than for TPR/FNR. Compared to the Non DP-Adam baseline, the FPR under DP-Adam is much higher, indicating more frequent false alarms at every $\epsilon$, where the TNR is significantly lower, reflecting weaker performance in correctly identifying true negatives (an individual who doesn't take medications or vitamins). This unexpected trend is unusual and highlights how applying DP can severely impact metrics like FPR and TNR, making model behavior more unstable.

# 6 Discussion and Conclusions

## 6.1 Part I: Private Aggregate Conclusion

Our experiments with differentially private aggregates and histograms on menstruation data demonstrate that basic DP mechanisms like the Laplace mechanism can produce useful, privacy-preserving statistics, even in relatively small datasets. For population-wide means and distributions, we observed that moderate privacy settings (e.g., $\epsilon \geq 1$) strike a reasonable balance between privacy and accuracy—providing meaningful insights while still adhering to strong privacy guarantees. These DP releases can allow users to compare their cycle lengths and menses data to broader population norms, potentially supporting personal health decisions and awareness without compromising individual privacy.

However, our results also show that stronger privacy settings (e.g., $\epsilon \leq 0.5$) result in high noise and less interpretable outputs—particularly problematic in health applications where outputs must be both accurate and intuitive. Moreover, stratifying results by demographics introduces additional trade-offs due to the division of the privacy budget, which can further exacerbate class imbalance issues and reduce accuracy.

While the private aggregate results are necessary for app users to get utility out of menstruation data, it is understandable to question the ethics of using DP to share private information with third-parties safely. We consider this a successful compromise because apps being able to sell data safely allows the app to be free and accessible for all. Further concerns could be eased with an opt-in method, where users can choose to have their data in the DP set sold to third-parties.

Overall, our findings suggest that DP can be effectively applied to the release of menstrual health statistics, especially in aggregate form, provided that privacy budgets are chosen carefully and datasets are sufficiently large. Future work could expand on this by combining these methods with synthetic data generation or personalized privacy budgets, to enhance both individual protection and data accuracy.

## 6.2 Part II: DP-SGD Model Conclusion

Our regression and classification experiments showcase the essential tradeoff between privacy and accuracy present in DP-SGD. As we imposed stronger privacy guarantees by decreasing the $\epsilon$ values, model performance significantly declined in terms of both regression, where we calculated the MSE and RMSE, and for classification. At higher $\epsilon$ values (e.g. 10 or more onward), we saw that it is possible to maintain a reasonable level of accuracy while offering moderate privacy protection to the data.

When applying DP to regression tasks involving sensitive health-related variables, careful tuning of the privacy parameter $\epsilon$ is essential. While stronger privacy guarantees are desirable from a data protection standpoint, they can severely limit model accuracy, especially for continuous outcome prediction. For real-world deployments, moderate privacy settings may offer a better privacy-accuracy tradeoff, balancing protection with reliable performance.

As for the binary classification task, in 7 we observed a clear decline in accuracy as we applied stronger privacy guarantees with lower $\epsilon$ values using DP-SGD with the Adam optimizer. The non-private Adam baseline significantly outperformed the DP-SGD Adam models, especially as $\epsilon$ dropped below 5, which highlights the difficulty of assigning stronger privacy guarantees at the cost of lower accuracy. Furthermore, we noticed a greater variability in test accuracy for smaller $\epsilon$ values, shown in the wider standard deviations, suggesting that lower privacy budgets not only reduce accuracy but also make model performance less stable across runs. This would be important for third parties to be aware of as inconsistent or inaccurate model outputs could lead to flawed decisions, especially in contexts like advertising or health related analytics.

Beyond overall accuracy, the additional metrics presented in 8 provided deeper insight into how differential privacy affects classification behavior. The TPR declined with smaller $\epsilon$ values, while FNR increased, indicating the models increasing difficulty in detecting individuals who take medications or vitamins under strong privacy guarantees where the baseline Adam model achieved a balanced middle ground. In contrast, FPR and TNR displayed less predictable trends with varied results across the $\epsilon$ values, but much worse performance compared to the DP Adam baseline which had a higher TNR and lower FPR. The irregular behavior suggests differential privacy could introduce instability specifically for the model handling negative cases as seen by the much higher TNR by the baseline Adam model.

For practitioners who are interested in applying differential privacy to sensitive health data, a key takeaway is the importance of aligning privacy settings with specific classification priorities of the application.

If the cost of missing true positives are high, then a lower $\epsilon$ value might be too costly in terms of accuracy. Conversely, if false positives are more problematic, the irregular behavior of FPR and TNR suggest using DP with caution, actually considering other methods rather than using DP seeing these poor results. These findings also suggest that when accuracy overall is more important for real world applications, stronger privacy guarantees may lead to poor tradeoffs, especially with low and unpredictable accuracy. Therefore, careful tuning of $\epsilon$ should be performed based on the desired privacy-accuracy balance, especially in high stakes situations like healthcare.

In our context, where data may be shared with third parties such as advertisers, inaccurate labeling could lead to flawed research conclusions or ethically questionable targeting. False positives are specifically more concerning in this setting. If a model incorrectly identifies someone as using medications or vitamins, advertisers could draw inaccurate conclusions about their health, violating contextual integrity and sending inappropriate advertisements to people who aren't actually taking medications or vitamins. This not only reduces the accuracy of the data for advertisers, but also raises serious ethical concerns around misrepresentation and stereotyping. Therefore, for applications involving third party data usage and advertising, minimizing false positives becomes an important priority. Applying DP made the false positive rate increase significantly and become more unstable, suggesting using DP may be too harmful for advertisers.

We hypothesize that the FPR was significantly higher for each $\epsilon$ value compared to the non-DP baseline because of the class imbalance and the nature of DP-SGD. Since the negative class (Individual does not take medications/vitamins) is the minority class, they already contribute fewer samples to the DP-SGD algorithm. The gradients from these rare and possibly more informative examples are clipped, which could decrease the learning feedback from these examples. Additionally, the noise added by DP can further skew the already sparse feedback from the minority class during training, hence the model becomes more biased to predicting the positive, majority class. So, the unstable FPR and TNR in our results could just be because of this imbalance, if more data was gathered and eliminated this imbalance it could provide more accurate insights into how DP-SGD affects this binary classification task.

### 6.3 Limitations

#### 6.3.1 Data

While our study provides valuable insights into the impacts of differential privacy on sensitive menstruation cycle data, several limitations must be discussed about the data itself. First, we had to clean the data very thoroughly to prepare for analysis as the original dataset was very messy. The dataset contained many missing values across important demographic features like age and ethnicity, but also menstrual cycle data like cycle length. Although data cleaning methods were used, these methods could make the data more biased and even more inaccurate, which in turn could harm results and conclusions.

Another limitation and problem we encountered was with class imbalance. Certain demographic groups in particular are underrepresented, which could skew analysis. Furthermore for the binary classification model there was class imbalance in the taking medications or vitamins variable, where more people in the dataset did use medications or vitamins. We used a weighted binary cross entropy to deal with the issue but it still could've harmed results and led to biased predictions for the majority group, which we see in the results could be the case. This imbalance makes results not as generalizable across different groups, reducing its reliability in the real world.

#### 6.3.2 Current Open-Source DP Libraries

From our experiments, we were quite surprised by the lack of proper state-of-the-art open-source DP libraries available currently. We first tried to do our experiments using the TensorFlow Privacy library (tf_privacy) but this module was slightly outdated and a bit more difficult to use, we spent quite a bit of time debugging and downgrading versions to get rather subpar results. While for this project, we have utilized Opacus by Meta, there is still rather limited functionality, and even some of the documentation isn't as helpful if we want to do proper research, as it provides very limited features. While DP is a very important topic in computer science and research involving data and individual privacy, there is a lot of great work done in the theoretical space but a lot more practical implementation is needed in the open-source space.

### 6.4 Future Work

There are several paths for future work that could build upon the findings of this paper and hopefully address some of the limitations mentioned earlier. One of the key challenges we faced was the handling of missing data. Future work could explore more advanced data cleaning techniques, or even better gather more clean and reliable data that could benefit analysis and make it more generalizable. While menstruation cycle data is limited online, studies could collect more data, hopefully in an anonymized way, that could be analyzed further. This could also fix the class imbalance problem, leading to a larger and more diverse dataset to better represent different groups.

While this study focused on applying specific DP techniques to this menstruation data, further research could also explore other privacy preserving techniques such as using synthetic data generation. Synthetic data generation could be used for multiple reasons, one could be to build a model using synthetic data rather than our approach with DP-SGD and observe those results, potentially with better accuracy and metrics results. Also, synthetic data could be used for releasing basic aggregate statistics with privacy guarantees, further enhancing the accuracy of privacy preserving analysis.

Finally, as privacy continues to be a major concern in healthcare specifically, future research could focus on privacy preserving techniques that can combine multiple methods to enhance model and aggregate performance, while ensuring strong privacy guarantees. As mentioned in the last paragraph, there are many more DP and non-DP privacy preserving methods to explore in the future, which could prove to be useful and beneficial in the healthcare industry and beyond. The findings from this paper can open the door for further advancements in privacy preserving techniques in the healthcare industry, which is an area that has had some progress but still offers a lot of room for improvement.

### 6.5 Contributions

Nathan: I primarily worked on the DP-SGD model building for the binary classification problem, and described the DP-SGD implementation in 4.2.1, explaining how the Opacus library works to apply differential privacy in stochastic gradient descent. I also wrote the methodology, feature selection, results, and conclusion for the binary classification problem. I also conducted background research and found much related work and wrote the background section alongside the abstract, as well as some of the limitations and future work. In terms of the poster I included my key results and visuals and restructured the content into concise bullet points.

Raima: Carried out the initial data cleaning, imputation, statistical, and exploratory analysis for feature selection to create the initial dataset for experimentation. I did the DP-SGD code experimentation for regression analysis using Opacus, from neural network building and feature selection to graphical result analysis. For the report, I wrote the DP-SGD Regression and general model parts in Methodology, Results, Conclusion/Discussion, and Open-source limitations. For the poster, I added the Introduction, Background, Data, DP-SGD Methodology diagram, Discussion, and Conclusion. My teammates did help in refining the write-ups and design orientation of the poster.

Lia: I performed initial research on advertisements, and concluded that DP-SGD would be a reasonable approach on behalf of advertisers needing to be able to target specific ads to users, while balancing user privacy. In the paper, I worked on the Python code for generating DP histograms, taking into account the max number of contributions for global sensitivity, etc. I contributed to the methodology as well, and I also created the 95% confidence intervals for the visualizations, discussed the privacy-accuracy trade-off, possible use cases of these statistics, and the conclusion for aggregate statistics. For the poster, I added conclusions for aggregate statistics, as well as the flowchart and visuals for the aggregate methods.

Shibani: Since this project was from my proposal, I took on the task of finding a useful dataset. I also carried out the final data deciphering and cleaning after we tested Raima's methods of imputation, because the original data had little documentation. I came up with the initial use cases of DP on menstruation data that we narrowed down to the two in this paper. For the code, I worked on displaying results with the functions as well as writing the code to create DP Means and DP histograms stratified by demographics at different values of epsilon. Although we had to narrow down the variables in the paper, the code was written for other menstruation data as well. Finally, I wrote parts of the paper including the introduction, parts of the histogram results & methodology, and the ethics discussions.

As for sharing our report, we choose option 3. share publicly on the course website.

# 7 Appendix

Table 2: Menses Length Means

| | True | $\epsilon = 0.1$ | $\epsilon = 0.5$ | $\epsilon = 1$ | $\epsilon = 5$ |
|---|---|---|---|---|---|
| Mean | 5.24 | 3.21 | 5.00 | 5.49 | 5.45 |

Table 3: Bleeding Intensity Means

| | True | $\epsilon = 0.1$ | $\epsilon = 0.5$ | $\epsilon = 1$ | $\epsilon = 5$ |
|---|---|---|---|---|---|
| Mean | 9.87 | 6.43 | 9.45 | 10.29 | 10.23 |

Figure 9: Menses Length Means by Demographics

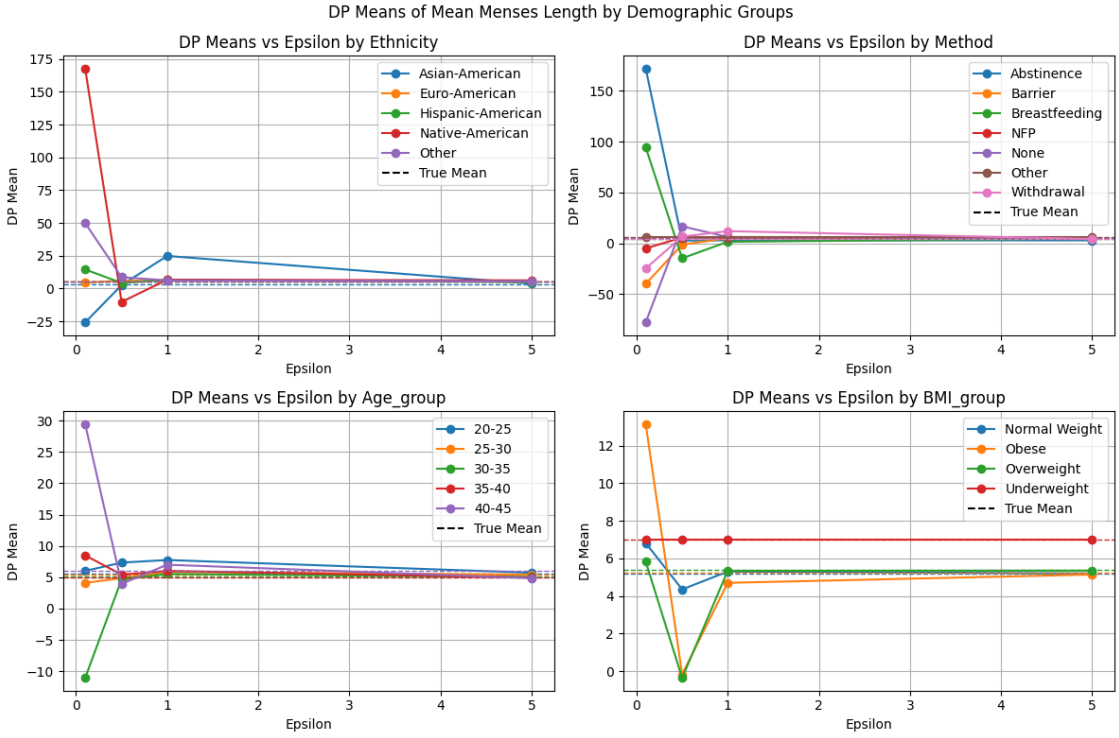


Figure 10: Bleeding Intensity Means by Demographics

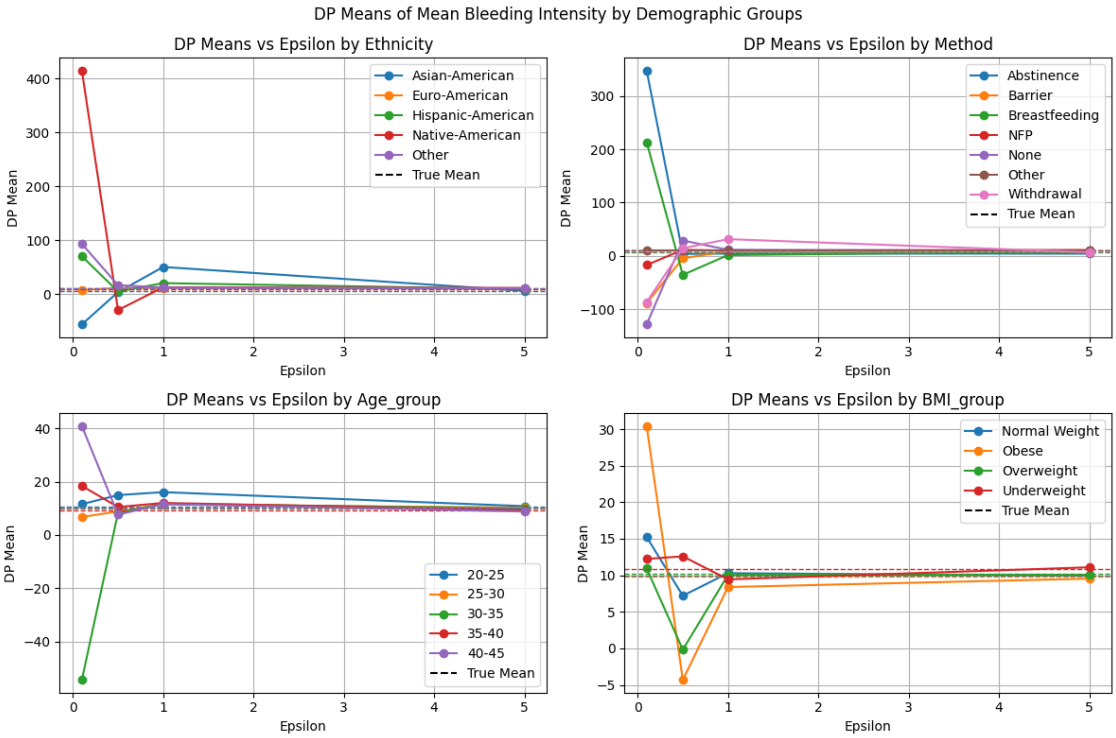



Figure 11: Mean Menses Length Distributions

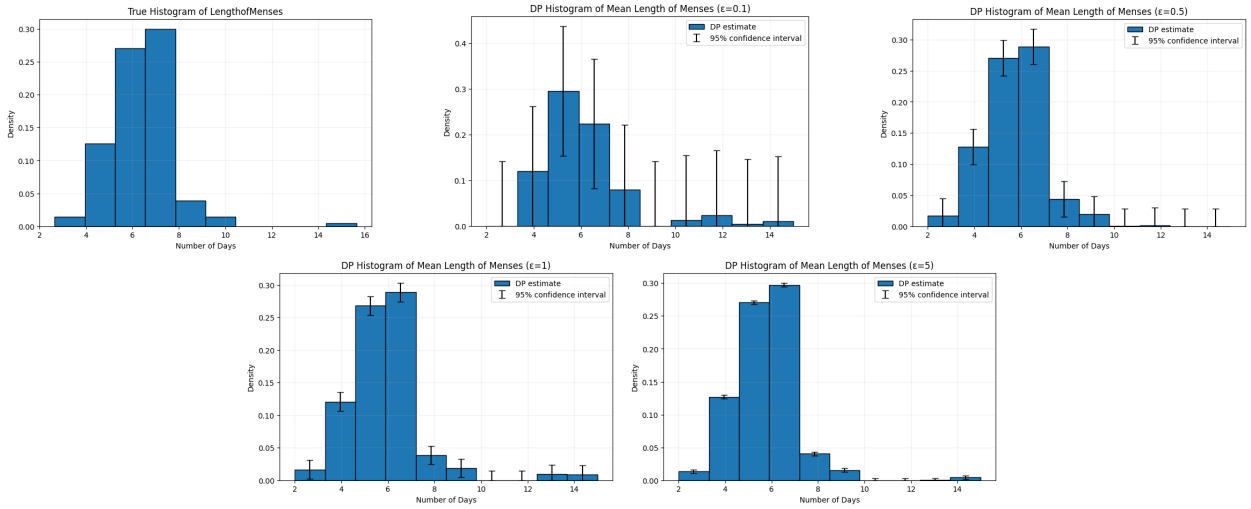


Figure 12: Menses Length Standard Deviation Distributions

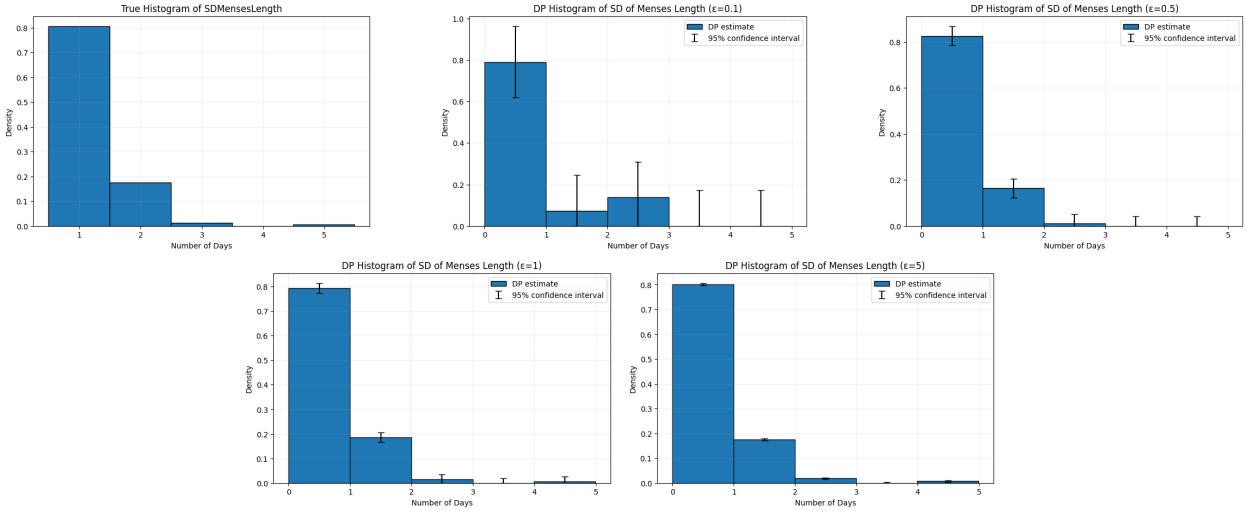
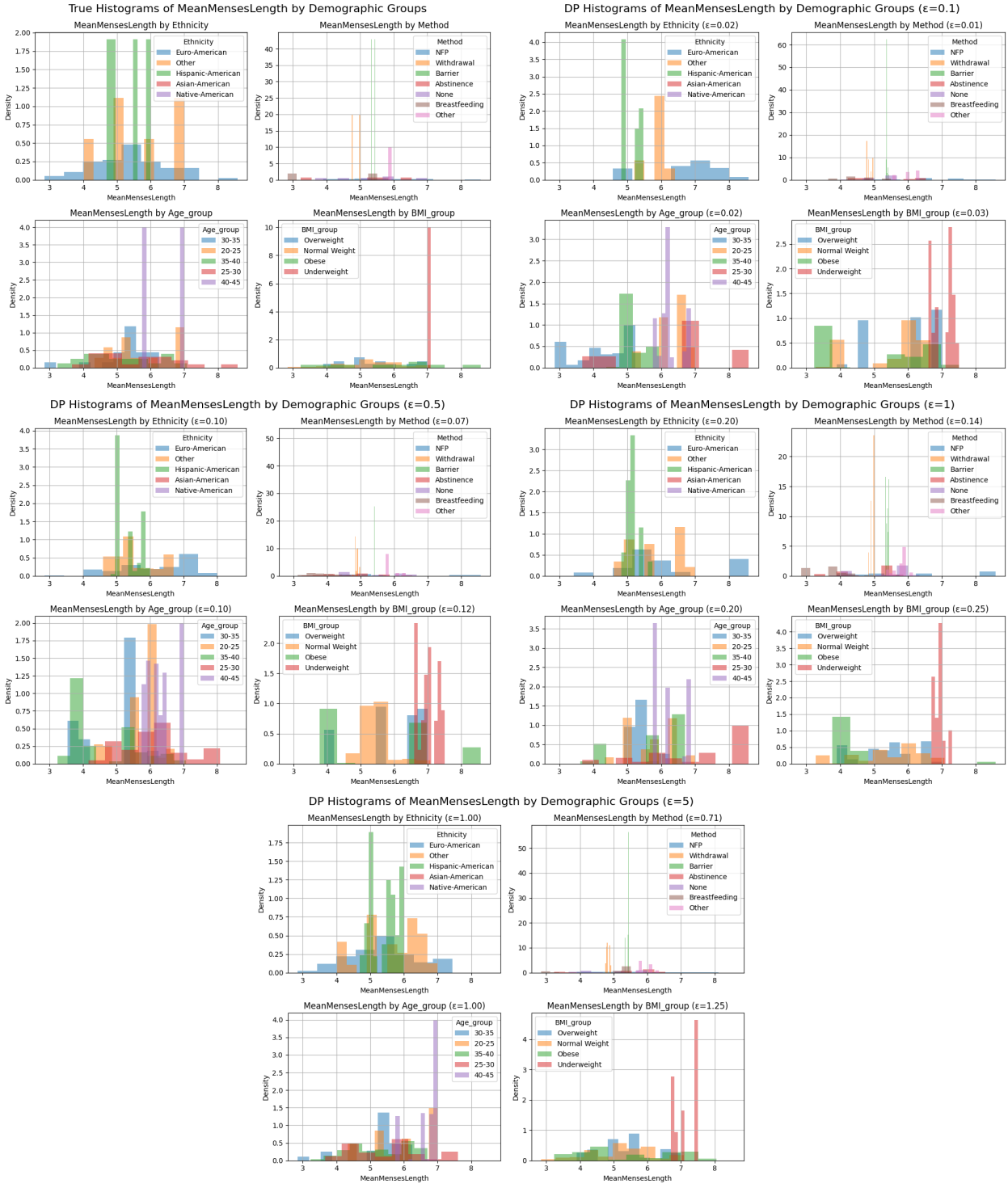


Figure 13: Menses Length Distributions by Demographics



## References


[1] Ghalia M Attia, Ohood A Alharbi, and Reema M Aljohani. "The Impact of Irregular Menstruation on Health: A Review of the Literature". In: *PMC* 2023 (2023). Accessed: 2025-04-25. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10733621/#:~:text=Menstrual%20irregularity%20has%20been%20found,their%20impact%20on%20women's%20health..

[2] Cleveland Clinic. *Ovulation: What Is It, Symptoms and Timing*. Accessed: 2025-04-26. 2023. URL: https://my.clevelandclinic.org/health/articles/23439-ovulation.

[3] A. Dyda et al. "Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality". In: *Patterns (N Y)* 2.12 (2021), p. 100366. DOI: 10.1016/j.patter.2021.100366.

[4] Richard J Fehring. "Menstrual Cycle Data". In: *Randomized Comparison of Two Internet-Supported Methods of Natural Family Planning* (2012). URL: https://epublications.marquette.edu/data_nfp/7.

[5] Simson Garfinkel. *Differential Privacy and the 2020 US Census*. Accessed: 2025-04-22. 2022. URL: https://mit-serc.pubpub.org/pub/differential-privacy-2020-us-census/release/2.

[6] Harvard T.H. Chan School of Public Health. *Menstrual Cycles Today: How Menstrual Cycles Vary by Age, Weight, Race, and Ethnicity*. Apple Women's Health Study Update. July 2024. URL: https://hsph.harvard.edu/research/apple-womens-health-study/study-updates/menstrual-cycles-today-how-menstrual-cycles-vary-by-age-weight-race-and-ethnicity/.

[7] Khalida Itriyeva. "The effects of obesity on the menstrual cycle". In: *Current Problems in Pediatric and Adolescent Health Care* 52.8 (2022), p. 101241.

[8] Bargav Jayaraman and David Evans. "Evaluating Differentially Private Machine Learning in Practice". In: *arXiv preprint arXiv:1902.08874* (2019). URL: https://arxiv.org/abs/1902.08874.

[9] Joel E. Kimmons et al. "Multivitamin Use in Relation to Self-Reported Body Mass Index and Weight Loss Attempts". In: *Journal of Nutrition* 137.5 (2007), pp. 1351–1357. URL: https://pubmed.ncbi.nlm.nih.gov/17406146/.

[10] W. Liu et al. "A Survey on Differential Privacy for Medical Data Analysis". In: *Ann Data Sci* (2023), pp. 1–15. DOI: 10.1007/s40745-023-00475-3.

[11] Mohammad Malekzadeh et al. *Dopamine: Differentially Private Federated Learning on Medical Data*. 2021. arXiv: 2101.11693 [cs.LG]. URL: https://arxiv.org/abs/2101.11693.

[12] BARRY M Sherman, STANLEY G Korenman, et al. "Hormonal characteristics of the human menstrual cycle throughout reproductive life." In: *The Journal of clinical investigation* 55.4 (1975), pp. 699–706.

[13] Medical News Today. *Anemia and periods: What you need to know*. Accessed: 2025-04-26. 2023. URL: https://www.medicalnewstoday.com/articles/anemia-and-periods.

[14] Dona Varghese et al. *Polypharmacy*. Accessed: 2025-04-26. Last Update: February 12, 2024. 2024. URL: https://www.continuingeducation.com/polypharmacy.