

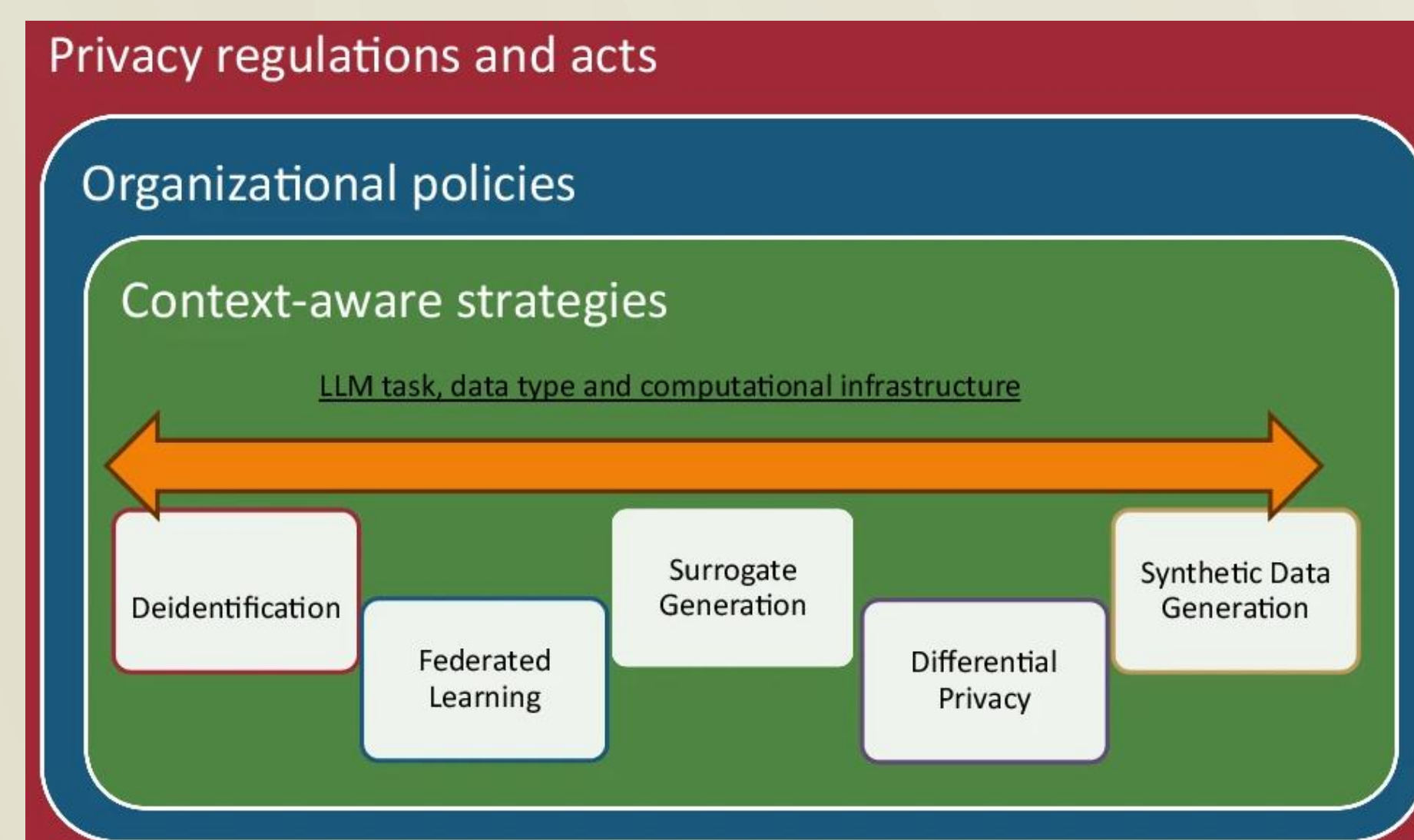
# Privacy-Preserving Methods for Oncology Research using Differential Privacy

Elena Valkanova, ValenaQuant

## INTRODUCTION

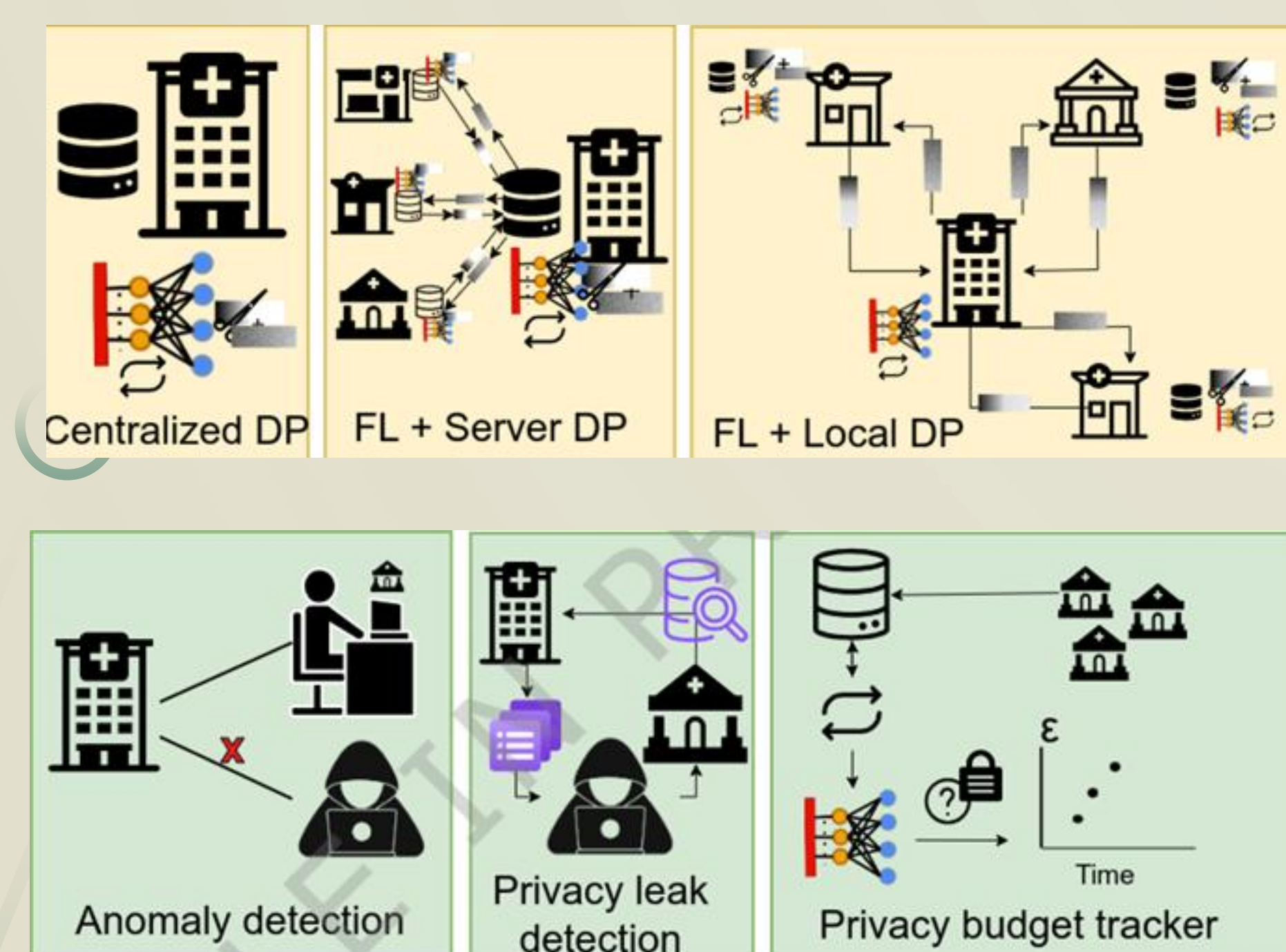
Differential Privacy (DP) enables analysis of sensitive data with formal, quantifiable privacy guarantees. This poster summarizes and compares major DP definitions (pure, approximate, Rényi, and zero-concentrated DP) and common mechanisms (Laplace, Gaussian, and Exponential), focusing on privacy–utility trade-offs and practical feasibility for oncology applications. We evaluated differential privacy (DP) methods using both real-world oncology data and simulated data designed to reflect common challenges, including small sample sizes, high dimensionality, complex interactions, and censoring.

Figure 1. Context-aware privacy preserving strategies



Jonnagaddala, J., Wong, Z.SY. Privacy preserving strategies for electronic health records in the era of large language models. *npj Digit. Med.* 8, 34 (2025). <https://doi.org/10.1038/s41746-025-01429-0>

Figure 2. Integrating (DP) into the lifecycle of medical learning models



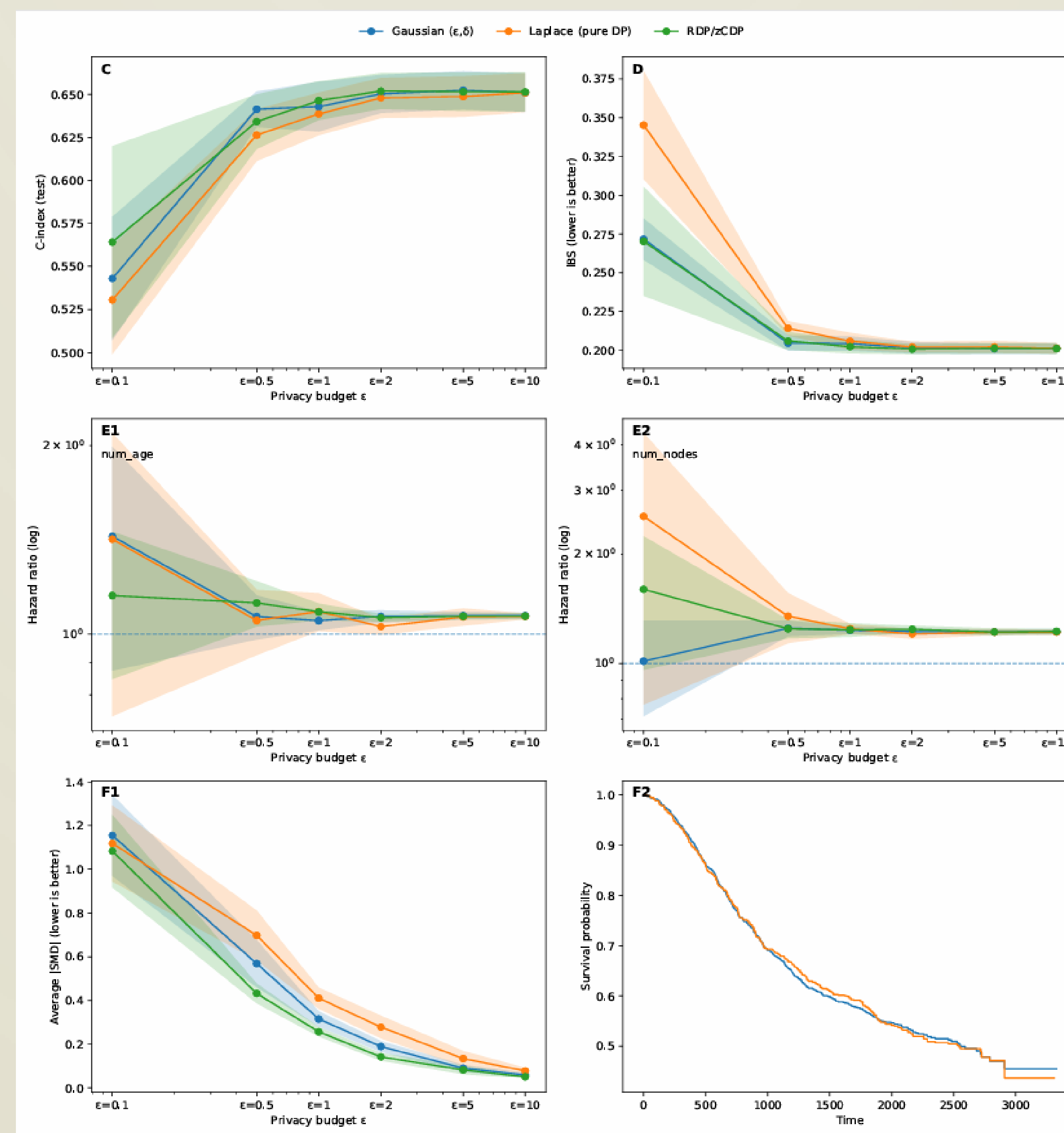
Mohammadi, M., Vejdanihemmat, M., Lotfinia, M. et al. Differential privacy for medical deep learning: methods, tradeoffs, and deployment implications. *npj Digit. Med.* (2026). <https://doi.org/10.1038/s41746-025-02280-z>

## DEFINITIONS

Table 1. Major DP definitions

| Definition                  | Formal statement (typical form)                                                                                                                                                          | Key parameter(s)   | Application                                                                                     |
|-----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|-------------------------------------------------------------------------------------------------|
| Pure DP                     | A randomized mechanism $M$ is $\epsilon$ -DP if for all neighboring datasets $D, D'$ and all outputs $S$ :<br>$P[M(D) \text{ in } S] \leq \exp(\epsilon) \cdot P[M(D') \text{ in } S]$ . | $\epsilon$         | Strong, clean guarantee; often conservative (more noise). Common for simple counts/queries.     |
| Approximate DP              | $(\epsilon, \delta)$ -DP if:<br>$P[M(D) \text{ in } S] \leq \exp(\epsilon) \cdot P[M(D') \text{ in } S] + \delta$ .                                                                      | $\epsilon, \delta$ | Allows a small failure probability $\delta$ ; enables better utility in high dimensions and ML. |
| Rényi DP (RDP)              | $(\alpha, \epsilon)$ -RDP if Rényi divergence of order $\alpha > 1$ satisfies:<br>$D_{\alpha}(M(D) \parallel M(D')) \leq \epsilon$ .                                                     | $\alpha, \epsilon$ | Convenient for composition/accounting (e.g., DP-SGD); converts to $(\epsilon, \delta)$ -DP.     |
| Zero-concentrated DP (zCDP) | $\rho$ -zCDP if for all $\alpha > 1$ :<br>$D_{\alpha}(M(D) \parallel M(D')) \leq \rho \cdot \alpha$ .                                                                                    | $\rho$             | Simple composition; closely tied to Gaussian noise; converts to $(\epsilon, \delta)$ -DP.       |

Figure 3. DP Benchmarking on Colon Survival Data: C-index, IBS, HR Stability, and Synthetic Fidelity



RDP = Rényi DP and zCDP = zero-concentrated DP; both provide convenient privacy accounting and can be converted to  $(\epsilon, \delta)$ -DP.

## METHODS

### DP benchmarking

We compared major DP definitions—pure DP, approximate DP, Rényi DP (RDP), and zero-concentrated DP (zCDP)—by implementing representative mechanisms and analysis workflows and then assessing privacy–utility trade-offs under matched privacy budgets.

For each method, we evaluated a grid of privacy budgets (e.g.,  $\epsilon \in \{0.1, 0.5, 1, 2, 5, 10\}$ ) and set  $\delta$  relative to sample size ( $\delta \approx 1/N^{(1+\gamma)}$ ) to reflect realistic disclosure risk in biomedical datasets.

### Utility and feasibility endpoints

We evaluated each DP approach across three dimensions.

- Predictive utility (primary): survival discrimination using Harrell's C-index on held-out test data; calibration/overall error using integrated Brier score (IBS).
- Statistical validity (secondary): bias/variance in key estimates and consistency of clinically meaningful associations (direction and relative magnitude).
- Implementation feasibility (practical): sensitivity to tuning (clip norms, noise multipliers, and accounting choices).

## SUMMARY

In our benchmarking framework, we compared core DP definitions and mechanisms under matched privacy budgets ( $\epsilon$ ). Using the colon cancer survival dataset (SurvSet\*), we quantified how increasing privacy protection affects (1) predictive discrimination (C-index), (2) calibration/error (integrated Brier score, IBS), and (3) interpretability through stability of key hazard ratios (e.g., age, node burden, treatment). To support safe data sharing, we also assessed synthetic data fidelity across  $\epsilon$  using distributional similarity metrics (average  $\|SMD\|$ ) and survival curve agreement (real vs synthetic Kaplan–Meier). Overall, Fig. 3 shows the expected privacy–utility trade-off: stronger privacy (smaller  $\epsilon$ ) is associated with lower predictive performance (C-index), higher error (IBS), and reduced synthetic-data fidelity (higher  $\|SMD\|$ ), with greater uncertainty in estimates; moderate  $\epsilon$  values better preserve clinically meaningful performance and effect stability, helping guide selection of practical privacy budgets.

\* Drysdale, E. (2022). SurvSet: An open-source time-to-event dataset repository. *ArXiv*, abs/2203.03094.

Your comments and questions are valued and encouraged.

evalkanova@valenaquant.com