Paper PP01

Privacy-Preserving Methods for Oncology Research Using Differential Privacy

Elena Valkanova, ValenaQuant, Tampa, US

ABSTRACT

Differential privacy (DP) provides formal, quantifiable guarantees that limit how much any single individual record can influence released statistics or trained models. We benchmarked representative DP definitions (pure DP, approximate (ε,δ)-DP, Rényi DP, and zero-concentrated DP) and common mechanisms (Laplace, Gaussian, and Exponential) on oncology survival settings. Using a real-world right-censored colon cancer dataset formatted in SurvSet and oncology simulated datasets, we evaluated privacy–utility trade-offs under matched privacy budgets ($\varepsilon \in \{0.1, 0.5, 1, 2, 5, 10\}$ with δ scaled to sample size). Utility was assessed via predictive discrimination (Harrell's C-index), overall error (integrated Brier score, IBS), stability of key hazard ratios, and synthetic data fidelity (standardized mean differences and Kaplan–Meier agreement). Results show the expected privacy–utility relationship: stronger privacy (smaller ε) increases uncertainty and degrades utility, while moderate ε values better preserve clinically meaningful performance and effect stability, supporting practical selection of privacy-preserving levels for biomedical analyses (Drysdale, 2022).

Keywords: differential privacy; survival analysis; oncology; synthetic data; privacy accounting

INTRODUCTION

Medical research increasingly relies on multi-site real-world data (RWD) and advanced machine learning but sharing patient-level information creates re-identification and inference risks. Traditional de-identification and anonymization approaches can reduce linking attacks yet may be vulnerable to auxiliary/background knowledge and model-based privacy attacks. Differential privacy (DP) addresses these limitations by bounding the change in an algorithm's output distribution when a single individual's record is added or removed, enabling transparent privacy–utility trade-offs and principled composition across repeated queries or iterative training (Dwork, 2006) (Dwork & Roth, 2014).

Recent work highlights the growing importance of privacy-preserving analytics for modern biomedical workflows, including electronic health records and large language models, as well as medical deep learning pipelines where membership inference and related attacks are well documented (Jonnagaddala & Wong, 2025) (Mohammadi et al., 2026). In parallel, regulatory and policy interest in the use of RWD and real-world evidence (RWE) continues to expand, including updated FDA communications and guidance on the use of RWE for regulatory

decision-making. These developments motivate practical benchmarking frameworks that connect formal privacy guarantees to performance, interpretability, and feasibility in realistic oncology settings (U.S. Food and Drug Administration [FDA], 2023) (FDA, 2026).

METHODS

Datasets

We evaluated DP methods on (i) a real-world oncology survival dataset and (ii) simulated oncology datasets designed to stress common challenges (small samples, high dimensionality, complex interactions, and right censoring). The real-world dataset was the colon cancer survival dataset distributed through SurvSet, containing time-to-event outcomes (time, event) and clinical predictors including age (num_age), number of positive nodes (num_nodes), treatment (fac_rx), sex (fac_sex), and tumor/clinical factors (fac_differ, fac_obstruct, fac_perfor, fac_adhere, fac_node4)(Drysdale, 2022).

Preprocessing included range checks and missingness review, encoding of categorical variables (one-hot or ordinal when clinically justified), and standardization of continuous predictors for model training. Train/test splits were stratified by the event indicator to preserve the censoring distribution and repeated across multiple random seeds to quantify variability.

Differential privacy definitions and representative mechanisms

We compared four common DP definitions—pure DP, approximate (ε,δ)-DP, Rényi DP (RDP), and zero-concentrated DP (zCDP)—by implementing representative mechanisms and workflows and evaluating privacy–utility trade-offs under matched privacy budgets.

Table 1. Major differential privacy definitions used for benchmarking.

Definition	Typical guarantee	Key parameter(s)	Notes / typical use
Pure DP	$P[M(D)\in S] \le \exp(\varepsilon)\cdot P[M(D')\in S]$	ε	Strong guarantee; often conservative (more noise).
Approximate DP	$P[M(D)\in S] \le \exp(\varepsilon)\cdot P[M(D')\in S] + \delta$	ε, δ	Allows small failure probability δ; useful in high dimensions / ML.
Rényi DP (RDP)	$D_\alpha(M(D)\|\|M(D')) \le \varepsilon_{RDP}$	α, εRDP	Convenient accounting/composition; convertible to (ε,δ)-DP.
zCDP	$D_\alpha(M(D)\|\|M(D')) \le \rho\cdot\alpha$	ρ	Simple composition; closely tied to Gaussian noise; convertible to (ε,δ)-DP.

Mechanisms were selected to match the definitions above: Laplace noise for pure DP releases, Gaussian noise for approximate DP and zCDP-style accounting, and Exponential mechanism for

private selection tasks (e.g., selecting a model/hyperparameter under a utility function). For iterative procedures (e.g., DP-SGD or repeated releases), privacy accounting used native RDP or zCDP where appropriate, with standardized reporting by converting to an equivalent (ε,δ)-DP guarantee for interpretability. (Dwork, McSherry, Nissim, & Smith, 2006) (McSherry & Talwar, 2007)

Privacy budgets and accounting

Each approach was evaluated across a grid of privacy budgets $\varepsilon \in \{0.1, 0.5, 1, 2, 5, 10\}$. The δ parameter was set relative to sample size to reflect realistic disclosure risk in biomedical datasets (e.g., $\delta \approx 1/N^{(1+\gamma)}$ for small $\gamma > 0$). We reported results comparably across methods by translating RDP/zCDP to (ε,δ)-DP when needed, while retaining native accounting for accurate composition in iterative pipelines.

Evaluation endpoints

We evaluated each DP approach across three dimensions:

- Predictive utility (primary): Harrell's C-index on held-out test data; integrated Brier score (IBS) for overall prediction error.
- Statistical validity (secondary): stability of key hazard ratios (direction and relative magnitude) for clinically meaningful covariates.
- Synthetic fidelity (data sharing): covariate balance via average absolute standardized mean differences (|SMD|) and survival curve agreement (real vs synthetic Kaplan–Meier).
- Implementation feasibility (practical): sensitivity to tuning choices (clipping, noise multipliers, accounting) and computational burden.

Simulated oncology datasets

Purpose and design principles

To complement the real-world SurvSet colon cancer dataset, we generated simulated oncology-time-to-event datasets to systematically evaluate DP behavior under controlled conditions. Simulations were designed to reflect practical oncology challenges that influence privacy–utility trade-offs, including small sample sizes, high dimensionality, correlated and interacting predictors, rare categories, and right censoring. Simulated scenarios allow us to vary these properties independently and quantify how privacy mechanisms affect prediction, calibration, effect stability, and synthetic fidelity as a function of the privacy budget ε.

Data-generating mechanism (time-to-event outcomes)

For each simulation replicate, we generated patient-level covariates X comprising a mix of continuous and categorical predictors, mirroring the structure of the colon dataset (e.g., age continuous variables, node-burden counts, and treatment/clinical categorical factors). Continuous predictors were generated from multivariate distributions with prespecified correlation structures (e.g., block correlation) to emulate realistic feature dependence. Categorical predictors were generated with prespecified marginal probabilities, including rare levels to mimic sparsity.

Event times T were generated using a proportional hazards framework, $\lambda(t \mid X) = \lambda_0(t) \cdot \exp(\eta(X))$, where $\lambda_0(t)$ is a baseline hazard (e.g., Weibull) and $\eta(X)$ is a linear predictor including main effects and selected interaction terms. Interactions were included to represent clinically plausible effect modification (e.g., treatment-by-risk). We additionally evaluated nonlinearity scenarios (e.g., threshold effects) to assess DP performance under mild model misspecification. Censoring times C were generated independently (administrative and/or random censoring), and observed outcomes were defined as $\tilde{T} = \min(T, C)$ with event indicator $\Delta = I(T \leq C)$. Censoring fractions were targeted to reflect oncology settings (e.g., 30%–70% censoring) and were varied across scenarios.

Simulation scenarios

- Sample size: $N \in \{200, 500, 1{,}000\}$ to represent small to moderate oncology cohorts.
- Dimensionality: $p \in \{20, 100, 500\}$ (after encoding) to capture low-, moderate-, and high-dimensional analyses.
- Signal strength: low/medium/high effect sizes to evaluate detectability under privacy noise.
- Correlation and collinearity: weak vs strong correlation blocks to reflect clinical feature redundancy.
- Rare strata: categorical levels with prevalence ~1%–5% to assess stability of subgroup estimates.
- Complexity: models with and without interactions/nonlinear terms to evaluate sensitivity to complex structure.
- Censoring: targeted censoring rates (e.g., 30%, 50%, 70%) to assess robustness under heavier censoring.

Alignment with DP benchmarking workflows

Each simulated dataset was processed using the same preprocessing and evaluation pipeline as the real-world analysis (encoding, standardization, stratified train/test splits, repeated seeds). DP definitions and mechanisms were benchmarked under the same privacy budget grid ($\varepsilon \in \{0.1, 0.5, 1, 2, 5, 10\}$ with δ scaled to sample size). RDP/zCDP accounting was used for iterative procedures, with standardized reporting converted to (ε,δ)-DP for comparability.

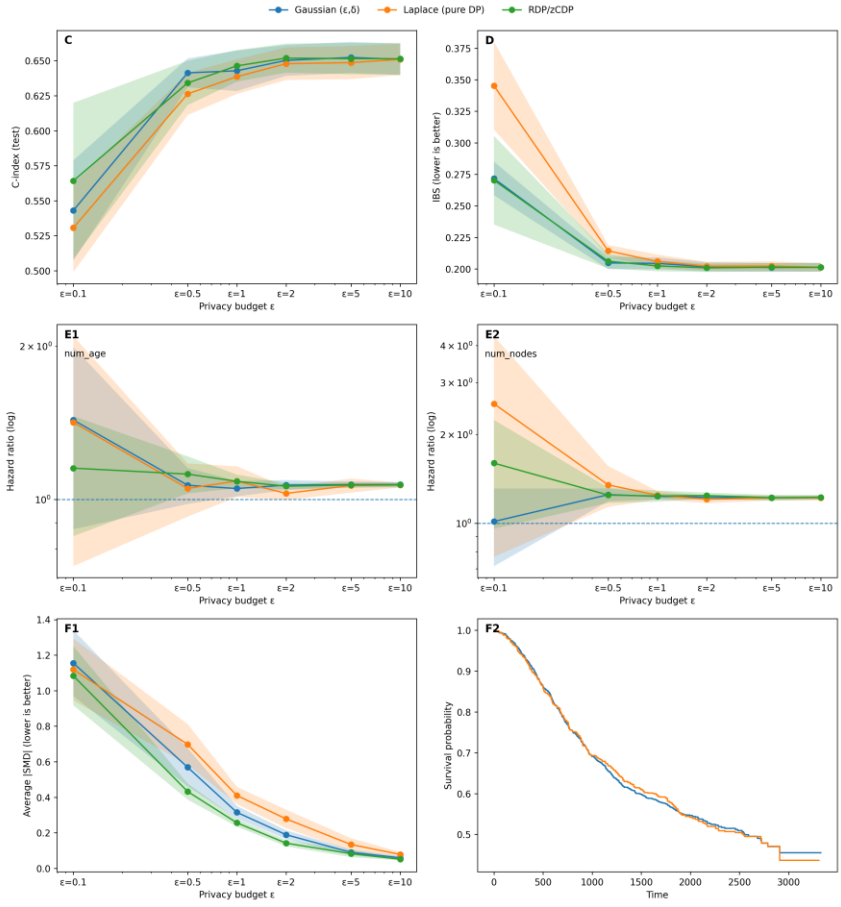
Evaluation endpoints (simulation)

- Predictive utility: C-index and IBS on held-out data.
- Effect recovery: bias and variability of key hazard ratios relative to known true coefficients (including selected interactions).
- Uncertainty and stability: across-seed variability and interval widening under stricter privacy.
- Synthetic fidelity: distribution and association preservation (e.g., average |SMD|, correlation agreement) and outcome fidelity (KM agreement).
- Failure modes: instability under rare strata, heavy censoring, and high dimensionality, recorded as convergence issues or large estimate dispersion.

RESULTS

Figure 1 summarizes DP benchmarking on the SurvSet colon dataset. As ε increased (weaker privacy), predictive discrimination improved (higher C-index), prediction error decreased (lower IBS), hazard ratio estimates stabilized, and synthetic data fidelity improved (lower |SMD|), consistent with an expected privacy–utility trade-off. Under strict privacy (small ε), uncertainty increased and utility degraded, particularly for effect estimation and synthetic fidelity (Drysdale, 2022). Simulation results were consistent with the real-data findings and provided additional insight into sample-size effects. Across all simulated scenarios, smaller privacy budgets ($\varepsilon \leq 0.5$) produced larger uncertainty and greater degradation in discrimination (lower C-index) and calibration/error (higher IBS), while moderate ε values (≈1–5) preserved clinically meaningful utility. The privacy impact was amplified in smaller samples (N=200), where hazard ratio estimates were less stable and synthetic fidelity metrics (|SMD| and Kaplan–Meier agreement) deteriorated more rapidly at strict privacy levels. As sample size increased (N=500 and N=1000), performance curves became smoother and effect estimates stabilized at lower ε, reflecting improved signal-to-noise robustness under privacy.

Figure 1. DP benchmarking on the SurvSet colon survival dataset: C-index vs ε, IBS vs ε, hazard ratio stability vs ε, and synthetic fidelity diagnostics



CONCLUSION

Our results illustrate how privacy budgets translate into practical performance for oncology survival modeling across both real-world and simulated settings. Across metrics with different orientations (higher-is-better vs lower-is-better), stronger privacy (smaller ε) produced larger perturbations, wider uncertainty bands, and reduced downstream utility, whereas moderate privacy budgets better preserved clinically meaningful discrimination and effect stability. In the SurvSet colon dataset, this pattern was observed consistently across C-index, IBS, and hazard ratio stability for key covariates, with synthetic fidelity diagnostics (average |SMD| and Kaplan-Meier agreement) supporting the same conclusion. Simulation experiments reinforced these findings and clarified the role of sample size. In smaller cohorts (e.g., N=200), utility degraded more quickly at strict privacy levels ($\varepsilon \leq 0.5$), and hazard ratio estimates exhibited greater dispersion, reflecting reduced signal-to-noise robustness under privacy noise. As sample size increased (N=500 and N=1000), privacy-utility curves became smoother and parameter stability improved at lower ε, suggesting that larger studies can tolerate stricter privacy while maintaining acceptable performance. This has direct implications for oncology RWD programs, where cohort sizes vary widely and pooling across sites may be necessary to support both privacy and utility. Taken together, the real-data and simulation results suggest a pragmatic workflow for selecting privacy budgets. First, choose candidate ε values based on the intended use: internal model development and decision support may tolerate lower ε than public release, whereas external sharing (including synthetic data release) typically benefits from conservative choices. Second, evaluate a small grid of ε values (e.g., 0.1-10) and report multiple endpoints: discrimination (C-index), calibration/error (IBS), and interpretability (stability of key hazard ratios). Third, quantify uncertainty across repeated splits/seeds and, for synthetic data, jointly assess covariate fidelity (|SMD| or related balance measures) and outcome fidelity (Kaplan-Meier agreement) to avoid over-reliance on any single summary. In practice, reporting should include the DP definition used, the mechanism (Laplace/Gaussian/Exponential), the sensitivity or clipping strategy, the privacy accountant and composition method, and the final (ε,δ) guarantee for the full pipeline. Where possible, hyperparameter selection should be performed with privacy-aware procedures (e.g., private model selection) to reduce the risk of privacy leakage through repeated tuning. Finally, these results highlight domain-specific challenges for oncology survival analysis under privacy constraints. High censoring, rare strata, and correlated covariates can amplify instability at strict privacy levels, particularly in small samples. For such settings, strategies that reduce sensitivity (robust preprocessing, clipping, regularization) and careful choice of evaluation metrics can materially improve practical performance.

LIMITATIONS AND FUTURE WORK

The findings are based on a single colon cancer dataset and representative implementations, so privacy–utility behavior may differ across diseases, endpoints, and modalities (e.g., imaging, genomics, longitudinal EHR) and under alternative DP survival models or robustness strategies. Although simulations help generalize trends and isolate sample-size effects, they still rely on

assumed data-generating mechanisms and may not capture all real-world complexities (missingness mechanisms, measurement error, site heterogeneity, and unmeasured confounding). Future extensions should benchmark additional time-to-event models (e.g., flexible hazards, competing risks) and assess federated implementations with secure aggregation and end-to-end privacy accounting, especially in multi-institution RWD environments.

DATA AND CODE AVAILABILITY

The real-world dataset used in this study is distributed through SurvSet (Drysdale, 2022). Analytical pipelines and figure-generation scripts are available upon request to support reproducibility and extension to additional datasets.

APPENDIX A. SIMULATION BENCHMARKING PANELS

Figures A1–A3 show the same DP benchmarking panel as Figure 1, evaluated on simulated oncology survival datasets at varying sample sizes (N=200, 500, 1000). Each panel reports C-index vs ε, IBS vs ε, hazard ratio stability for age and node burden, and synthetic fidelity diagnostics (|SMD| vs ε and Kaplan–Meier agreement at ε=1).

Figure A1. DP benchmarking panel on simulated oncology survival data (N=200).

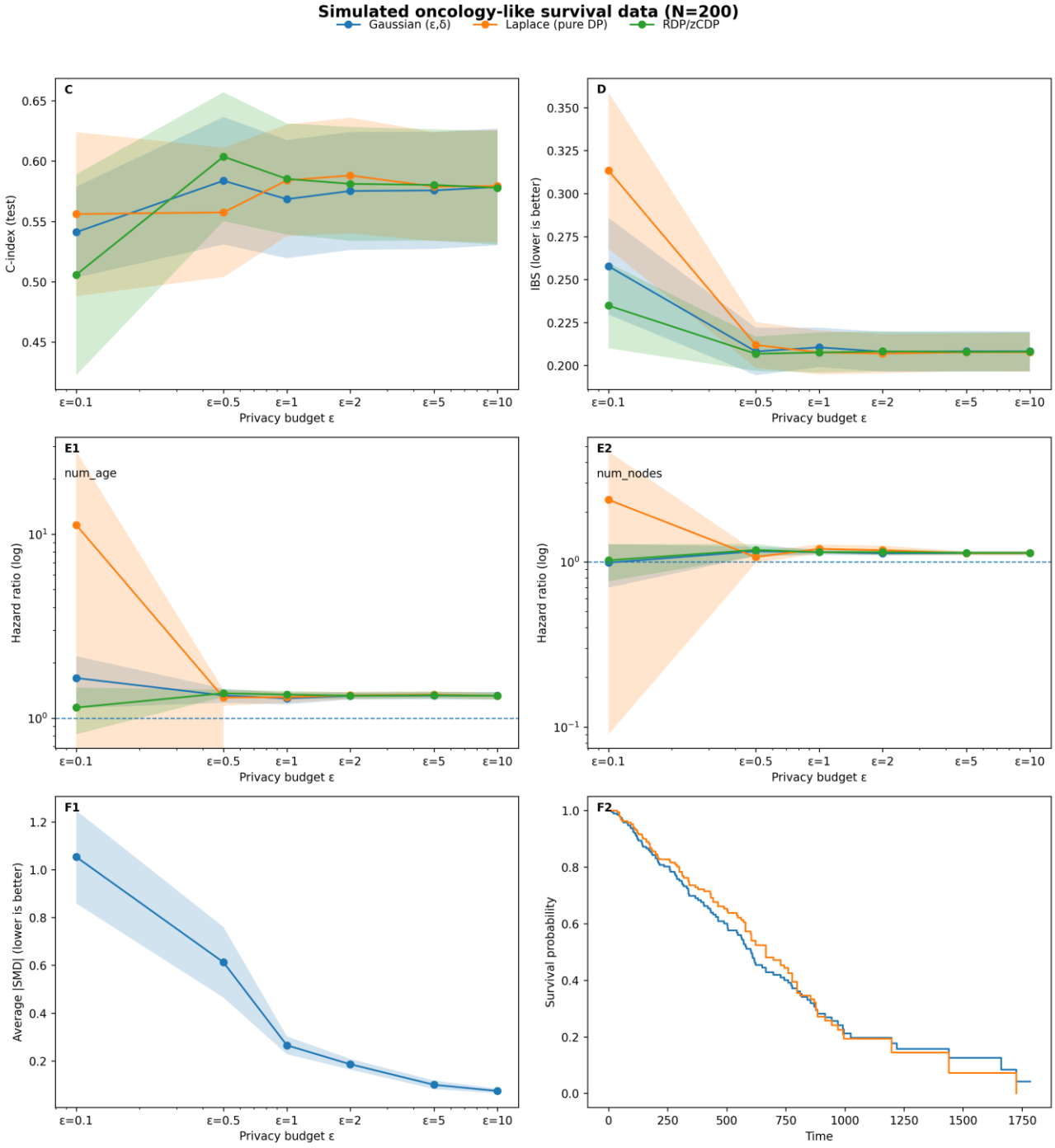


Figure A2. DP benchmarking panel on simulated oncology survival data (N=500).

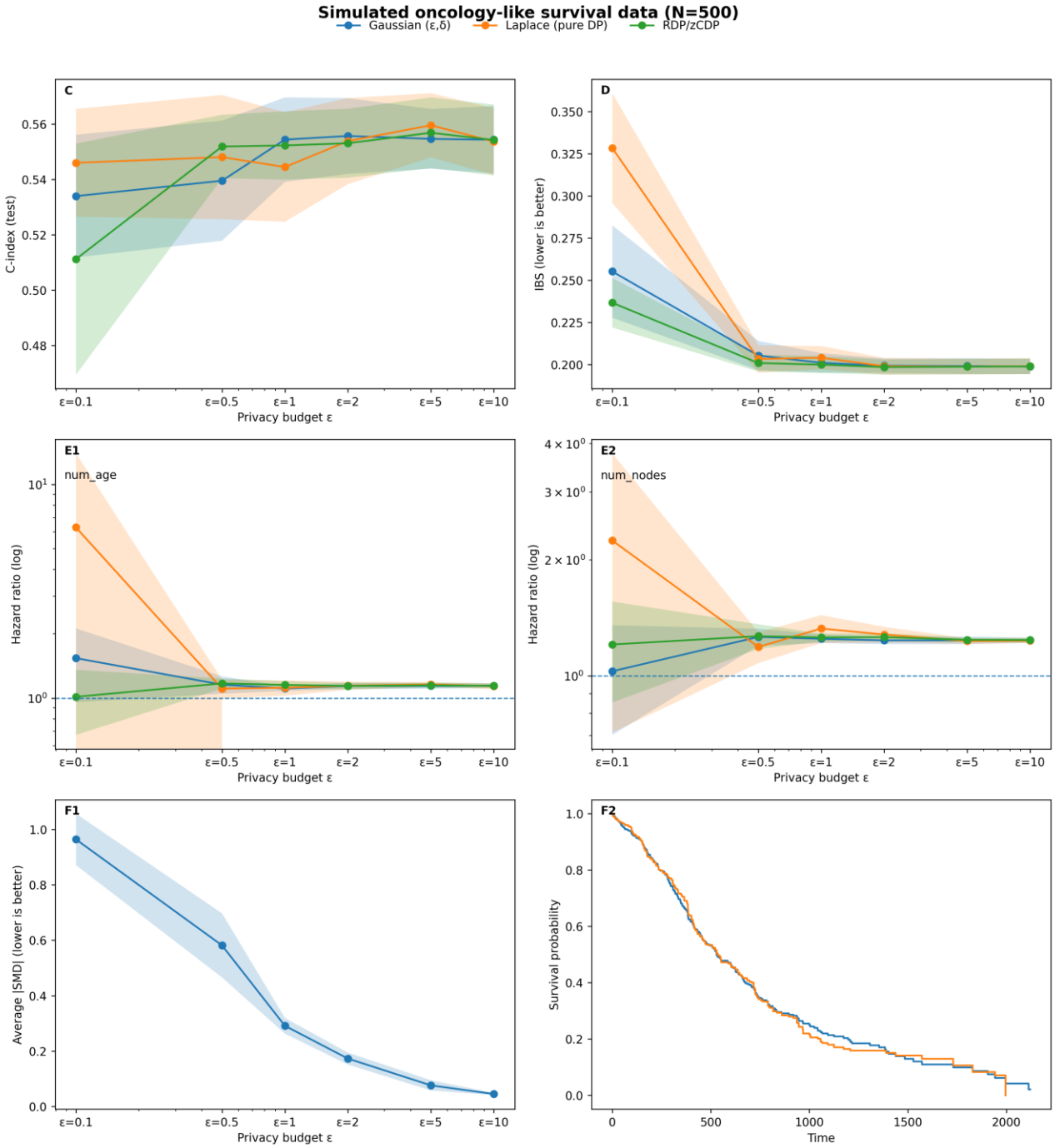
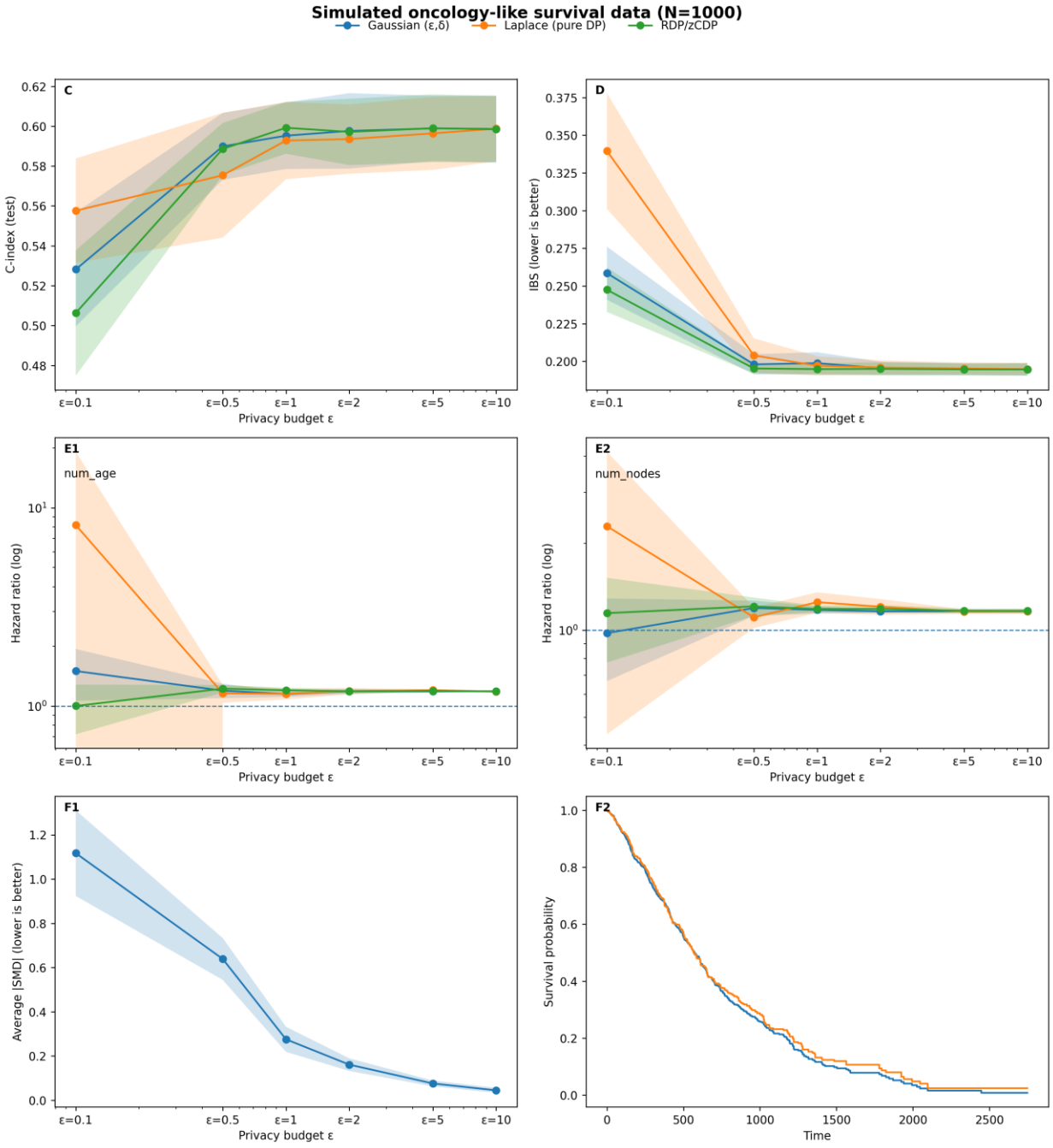


Figure A3. DP benchmarking panel on simulated oncology survival data (N=1000).



REFERENCES

Drysdale, E. (2022). SurvSet: An open-source time-to-event dataset repository. arXiv. https://arxiv.org/abs/2203.03094

Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener (Eds.), Automata, languages and programming (pp. 1–12). Springer. https://doi.org/10.1007/11787006_1

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Journal of Privacy and Confidentiality, 7(3), 17–51.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

Jonnagaddala, J., & Wong, Z. S.-Y. (2025). Privacy preserving strategies for electronic health records in the era of large language models. npj Digital Medicine, 8, Article 34. https://doi.org/10.1038/s41746-025-01429-0

McSherry, F., & Talwar, K. (2007). Mechanism design via differential privacy. Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 94–103. https://doi.org/10.1109/FOCS.2007.66

Mohammadi, M., Vejdanihemmat, M., Lotfinia, M., Rusu, M., Truhn, D., Maier, A., & Tayebi Arasteh, S. (2026). Differential privacy for medical deep learning: methods, tradeoffs, and deployment implications. npj Digital Medicine, 9, Article 93. https://doi.org/10.1038/s41746-025-02280-z

U.S. Food and Drug Administration. (2023, August 30). Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drug and biological products (Guidance for industry).

U.S. Food and Drug Administration. (2024, July 25). Real-world data: Assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products (Guidance for industry).

U.S. Food and Drug Administration. (2025, December 17). Use of real-world evidence to support regulatory decision-making for medical devices (Guidance for industry and Food and Drug Administration staff).

U.S. Food and Drug Administration. (2026, January 23). Real-world evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

CONTACT INFORMATION:

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Elena Valkanova

Company: ValenaQuant, Inc

Email: evalkanova@valenaquant.com