“Dead or Alive, we can deny it”. A Differentially Private Approach to Survival Analysis.

SEBD 2024
32nd SYMPOSIUM ON ADVANCED DATABASE SYSTEMS

Authors:
Francesco L. De Faveri, *University of Padova*
Guglielmo Faggioli, *University of Padova*
Nicola Ferro, *University of Padova*
Riccardo Spizzo, *IRCCS CRO-Aviano*

June 23-26, 2024 - Villasimius, Sardinia, Italy

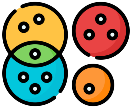
Università degli Studi di Padova

Table of Contents

Medical Analysis and Privacy

Introduction

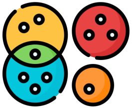
- Clustering patients based on similar characteristics, such as diseases or treatments, to obtain insights for the research.
- Researchers calculate the **survival probability** of new patients belonging to the population computed.



Medical Analysis and Privacy

Introduction

- Clustering patients based on similar characteristics, such as diseases or treatments, to obtain insights for the research.
- Researchers calculate the **survival probability** of new patients belonging to the population computed.



Privacy Risks

*The **more sensitive** the analysis is, the **higher the risk** of leaking sensitive patients' information grows.*

Problem Statement

Introduction

Can we apply ε-Differential Privacy and investigate the **Privacy** *vs.* **Utility trade-off** when performing Survival Analysis?

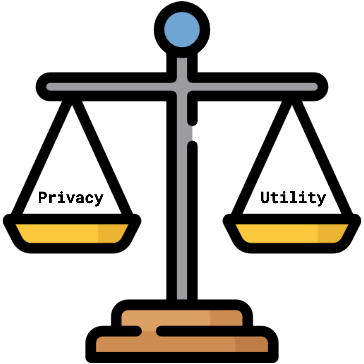


Table of Contents

Background & Related Works

Survival Analysis

Background & Related Works

Survival Analysis is a statistical technique used to analyze time-to-event or time-to-failure data when an event of interest has not yet occurred at time t.

Specifically, the Kaplan-Meier method [5] is used to evaluate the **survival trends** of the patients with common characteristics.

Kaplan-Meier Estimator:

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(\frac{n_i - d_i}{n_i} \right)$$

where:

d_i is the event occurrence.

n_i is the count of patients who have not experienced the event analysed at time $t_i \leq t$.

Definition (Dwork et al. [2])

A randomized mechanism $\mathcal{M}$, i.e., an algorithm that takes an input and returns a noisy output, is **ε-Differentially Private** iff for any pair of neighbouring datasets D and D', i.e., datasets that differ for at most one record, and a privacy budget $\varepsilon \in \mathbb{R}^+$, it holds:

$$\Pr\left[\mathcal{M}(D) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{M}(D') \in S\right] \quad \forall S \subset \mathrm{Im}(\mathcal{M})$$

Remark:

The lower the ε, the higher the privacy guarantees provided by the mechanism $\mathcal{M}$.

Privacy Loss

Background & Related Works

To verify that a mechanism $\mathcal{M}$ is ε-Differentially Private, the **Privacy Loss** of the mechanism must be bounded by the privacy budget ε with probability 1 [3].

$$\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}(O) = \log\left(\frac{\Pr\left[\mathcal{M}(D) = O\right]}{\Pr\left[\mathcal{M}(D') = O\right]}\right)$$

where:

D and D' are neighbouring datasets.

O is an output of the mechanism.

Related Works

The other proposed method (**LNTE**) to obfuscate Survival Analysis [4] was based on the Laplacian mechanism to change the number of subjects at risk and events in a dataset.

The algorithm iteratively **adjusts these counts** across time points to maintain updated risk and event information, returning a differentially private estimation of the survival rate.

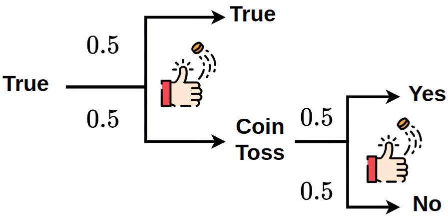
Table of Contents

Methodology

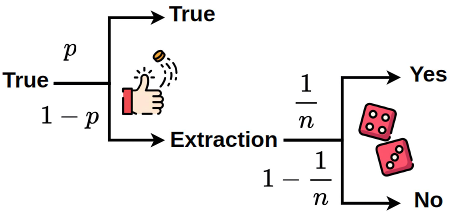
- Introduction
- Background & Related Works
- Methodology
- Experimental Results
- Conclusion

Revised Randomized Response (RRR)

Methodology



(a) Original Randomize Response mechanism.



(b) Revised Randomized Response mechanism.

Privacy Properties

These equations provide the probability of obtaining as the output the real category C_i, considering the input category C_i or another category $\overline{C_i}$.

$$\Pr[Resp = C_i | True = C_i] = p + (1-p)\frac{1}{n} = \frac{np + 1 - p}{n}$$

$$\Pr[Resp = C_i | True = \overline{C_i}] = (1-p)\left(1 - \frac{1}{n}\right) = \frac{n - 1 - np + p}{n}$$

Privacy Condition

We compute the Privacy Loss of the RRR mechanism:

$$\varepsilon(n, p) = \log\left(\frac{np + 1 - p}{n - 1 - np + p}\right)$$

Therefore, by **fixing** the number of categories n, the condition to satisfy the ε-Differential Privacy definition is:

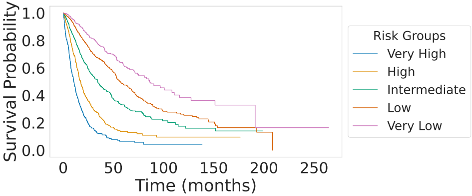
$$\varepsilon > 0 \iff p \in \left(\frac{n-2}{2n-2}, 1\right)$$

Table of Contents

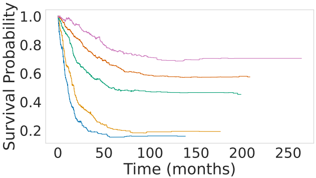
Experimental Results

Kaplan Meier Survival Curves: Original *vs.* LNTE

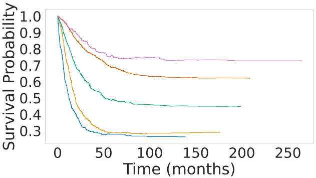
Experimental Results: IPSS-R Dataset [1]



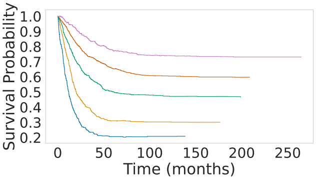
(c) IPSS-R



(d) LNTE, $\varepsilon = 1$



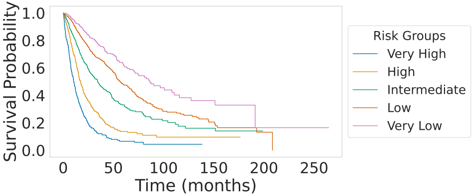
(e) LNTE, $\varepsilon = 2$



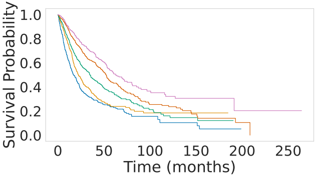
(f) LNTE, $\varepsilon = 3$

Kaplan-Meier Survival Curves: Original *vs.* RRR

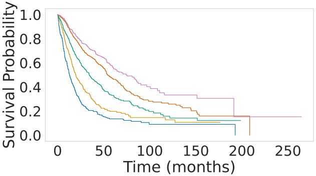
Experimental Results: IPSS-R Dataset [1]



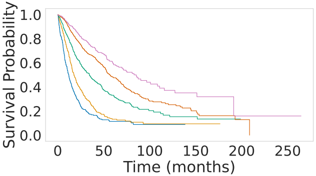
(g) IPSS-R



(h) RRR, $\varepsilon = 1$



(i) RRR, $\varepsilon = 2$



(j) RRR, $\varepsilon = 3$

Pairwise Log-Rank Test: Original *vs.* RRR

Experimental Results: Kidney Dataset [6]

Disease A	Disease B	Test Statistics		p-value		$-\log_2(p)$	
		Original	$\varepsilon = 3$	Original	$\varepsilon = 3$	Original	$\varepsilon = 3$
AN	GN	0.01	0.11	0.93	0.75	0.11	0.42
	Other	1.69	0.85	0.19	0.36	2.37	1.48
	PKD	1.09	0.79	0.30	0.37	1.75	1.42
GN	Other	0.99	0.63	0.32	0.43	1.64	1.23
	PKD	0.60	0.60	0.44	0.44	1.19	1.19
Other	PKD	0.26	0.39	0.61	0.53	0.71	0.91

Median Survival times

Experimental Results

Dataset	Mech.	Category	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	No DP
Kidney	LNTE	AN	1.30 (1.13, -)	1.30 (1.00, -)	1.33 (1.00, -)	1.77
		GN	1.00 (0.50, -)	4.33 (0.73, -)	5.13 (0.87, -)	1.00
		PKD	18.73 (5.07, -)	5.07 (1.00, -)	5.07 (2.10, -)	2.60
		Other	3.97 (1.80, -)	5.90 (2.10, -)	8.17 (3.80, -)	4.70
	RRR	AN	2.20 (1.27, 6.53)	1.43 (0.90, 3.20)	1.77 (0.90, 3.20)	1.77
		GN	0.93 (0.40, 4.33)	1.27 (0.50, 5.20)	1.30 (0.50, 5.20)	1.00
		PKD	2.60 (0.50, 9.73)	5.07 (0.87, 17.03)	4.40 (1.00, 5.07)	2.60
		Other	5.07 (0.80, 14.9)	3.97 (0.80, 9.73)	4.70 (0.93, 8.17)	4.70

Table of Contents

Conclusion

Contributions & Future Work

Conclusion

Our findings suggest that the RRR mechanism:

- Results in a more effective privacy and utility balance.
- Maintains the distribution properties of real results.
- Offers a new comparison method for future investigations.

As Future Directions:

- Investigate further the privacy guarantees offered by the Differential Privacy mechanism.
- Propose new privacy strategies for bridging the gap between Survival Analysis and Differential Privacy, e.g., Linear and Cox Regression with Differential Privacy.

“Dead or Alive, we can deny it”. A Differentially Private Approach to Survival Analysis.

Thank you for listening!

Any questions?

References

Backup-Slides

E. Bernard, H. Tuechler, P. L. Greenberg, R. P. Hasserjian, J. E. A. Ossa, Y. Nannya, S. M. Devlin, M. Creignou, P. Pinel, L. Monnier, G. Gundem, J. S. Medina-Martinez, D. Domenico, M. Jädersten, U. Germing, G. Sanz, A. A. van de Loosdrecht, O. Kosmider, M. Y. Follo, F. Thol, L. Zamora, R. F. Pinheiro, A. Pellagatti, H. K. Elias, D. Haase, C. Ganster, L. Ades, M. Tobiasson, L. Palomo, M. G. D. Porta, A. Takaori-Kondo, T. Ishikawa, S. Chiba, S. Kasahara, Y. Miyazaki, A. Viale, K. Huberman, P. Fenaux, M. Belickova, M. R. Savona, V. M. Klimek, F. P. S. Santos, J. Boultwood, I. Kotsianidis, V. Santini, F. Solé, U. Platzbecker, M. Heuser, P. Valent, K. Ohyashiki, C. Finelli, M. T. Voso, L.-Y. Shih, M. Fontenay, J. H. Jansen, J. Cervera, N. Gattermann, B. L. Ebert, R. Bejar, L. Malcovati, M. Cazzola, S. Ogawa, E. Hellström-Lindberg, and E. Papaemmanuil.
Molecular international prognostic scoring system for myelodysplastic syndromes.

References

Backup-Slides

NEJM Evidence, 1(7):EVIDoa2200008, 2022.

C. Dwork, F. McSherry, K. Nissim, and A. D. Smith.
Calibrating noise to sensitivity in private data analysis.
In S. Halevi and T. Rabin, editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

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

References

Backup-Slides

- L. Gondara and K. Wang.
  Differentially private survival function estimation.
  In F. Doshi-Velez, J. Fackler, K. Jung, D. C. Kale, R. Ranganath, B. C. Wallace, and J. Wiens, editors, *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2020, 7-8 August 2020, Virtual Event, Durham, NC, USA*, volume 126 of *Proceedings of Machine Learning Research*, pages 271–291. PMLR, 2020.

- E. L. Kaplan and P. Meier.
  Nonparametric estimation from incomplete observations.
  Journal of the American Statistical Association, 53(282):457–481, 1958.

References

Backup-Slides

C. A. McGilchrist and C. W. Aisbett.
Regression with frailty in survival analysis.
Biometrics, 47(2):461–466, 1991.