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

Francesco Luigi, De Faveri[1,*], Guglielmo, Faggioli[1], Nicola, Ferro[1] and Riccardo, Spizzo[2]

[1]*Department of Information Engineering, University of Padua, Padua, Italy*
[2]*National Cancer Center CRO Aviano, Aviano, Italy*

Abstract

Survival Analyses (SAs), a key statistical tool used to predict event occurrence over time, often involve sensitive information, necessitating robust privacy safeguards. This work demonstrates how the Revised Randomized Response (RRR) can be adapted to ensure Differential Privacy (DP) while performing SAs. This methodology seeks to safeguard the privacy of individuals' data without significantly changing the utility, represented by the statistical properties of the survival rates computed. Our findings show that integrating DP through RRR into SAs is both practical and effective, providing a significant step forward in the privacy-preserving analysis of sensitive time-to-event data. This study contributes to the field by offering a new comparison method to the current state-of-the-art used for SAs in medical research.

Keywords

Differential Privacy, Privacy-Preserving Mechanisms, Survival Analysis, Information Security

1. Introduction

Large amounts of data have been instrumental in medical research, leading to significant advancements and important scientific discoveries. Thanks to the availability of big data in healthcare [1, 2], researchers have improved their understanding of certain diseases and developed powerful prognostic models. A general method in medical data analysis involves clustering patients based on similar characteristics, such as diseases or treatments, to obtain insights for research purposes. Once the clusters are created, a critical aspect that researchers consider is understanding the survival probability of new patients belonging to a population. Researchers commonly resort to statistical procedures called Survival Analyses (SAs) to achieve this objective. This technique helps study the probability of survival for patients belonging to a specific population based on personal and sensitive data collected during clinical trials [3, 4, 5].

However, with the increasing reliance on data, particularly those containing personally identifiable information, there arises a significant concern regarding privacy [6, 7, 8]. Emerging from leaking sensitive information from survival research, malicious employers may decide to terminate employment before covering medical expenses incurred due to the condition if they know about the employee's medical condition. To avoid such a situation, the gold standard definition of privacy [9] has been introduced in the medical research domain. Differential

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23–26, 2024, Villasimius, Sardinia, Italy
*Corresponding author.
francescoluigi.defaveri@phd.unipd.it (F. L. De Faveri); faggioli@dei.unipd.it (G. Faggioli)
0009-0005-8968-9485 (F. L. De Faveri); 0000-0002-5070-2049 (G. Faggioli); 0000-0001-9219-6239 (N. Ferro); 0000-0001-7772-0960 (R. Spizzo)

Privacy (DP) provides the patient with *"Plausible Deniability"*, a condition under which an individual can deny his participation in a specific research study in a manner that an adversary cannot disprove with certainty. Applying Differential Privacy (DP) have been used over the last decade [10, 11, 12], yet without considering specifically the task of SAs.

In this study, we show how to use a revised version of the randomized response, i.e., the coin-toss mechanism, introduced by Warner [13] and modified by Greenberg et al. [14], in order to protect patients' sensitive categories when performing SAs for medical research purposes. We provide formal proof of the ε-DP property and report a comparison with a similar privatization mechanism used in literature with the same goal, proving that in DP scenarios, with an appropriate privacy budget ($\varepsilon = 3$) the results maintains important characteristics of the original results. The main contribution of this paper is to apply a DP mechanism in SAs and expose the trade-off between utility and privacy in performing SA in a differentially private manner using the coin-toss mechanism to achieve privacy for the patients.

In Section 2, we describe the related works used to provide privacy in healthcare research, focusing on privacy in SAs. Section 3 explains the theoretical background of the work, and Section 4 illustrates the mechanism used to address the problem of privacy in the SA function computation. Finally, in Section 5, we report our findings and fully discuss the results obtained.

2. Related Works

Privacy-preserving technology literature has offered different techniques to adopt DP in medical research studies, such as generative artificial intelligence and statistical models [15, 16].

Solutions for providing privacy using artificial intelligence methods include the implementation of Generative Adversarial Networks (GANs) [17, 18, 19], trained using DP, to create new synthetic data for research purposes rather than operating with the real data. However, GANs have limitations in terms of scalability and efficiency, which include the need for a significant amount of training data and higher demands on power consumption as reported in [20, 21].

On the other hand, the application of statistical models for survival analysis was first explored by Nguyên and Hui [22], Yu et al. [23], which investigated the impact of DP to understand the impact of explanatory variables for discrete-time SAs. In [22] the author proposed an extension of the DP Output Perturbation [24] method originally proposed for the empirical risk minimization problem. However, such an approach is limited for discrete-time SAs and is not applicable in this study. Moreover, Yu et al. [23] proposes to project patients' data to a space of lower dimension such that the projection preserves good characteristics of the original data. Nevertheless, the method does not rely on any formal definition of privacy.

To the best of our knowledge, it is not present in the literature a model to perform SA privatizing the categories in which patients are grouped. A similar work has been proposed by Gondara and Wang [25]. The authors proposed the Laplacian Noise Time Event (LNTE) mechanism to obfuscate the time-to-event data used by researchers when investigating the survival rates of a patients' dataset. The LNTE mechanism modifies traditional survival analysis by introducing Laplace noise, in line with a specified privacy budget ε, to both the subjects at risk r_j and events d_j within a dataset D. This process generates a perturbed matrix M', ensuring DP. Then, the algorithm iteratively adjusts these counts across time points to maintain

updated risk and event information, culminating in a differentially private estimation of the survival probability. However, the mechanism has some limitations, especially when the dataset size is small, leading to truncation and biased estimates in the survival rates.

3. Background and Preliminaries

We split the background into two parts: initially, we briefly describe SA and its main use in healthcare. Then present the DP framework, describing the properties used in our methodology.

3.1. Survival Analysis

Survival Analyses (SAs) are statistical techniques commonly used to analyze time-to-event or time-to-failure data when the event of interest has not yet occurred [26, 27]. Such functions investigate the time between a dichotomous event to occur and are used in different fields of research [28, 29]. In SAs, given a certain population of interest, researchers observe the expression of some event at a defined time t and want to compute the probability trend of some event, e.g., the occurrence of a certain symptom, the need of a certain treatment, or the death, for the rest of the population at time $t' > t$. Focusing on the medical research we are presenting, the event of interest is often defined as the *death* of a patient belonging to a specific population. An intuitive definition of the survival function is that it provides the probability that the event of interest, i.e., the death of a patient, has not yet occurred by time t. On the other hand, a *censored* event refers to a situation where the exact time of the event (such as death, relapse, or recovery) is unknown for an individual within the study. Censoring occurs when the observation period ends before the event occurs or when the individual is lost to follow-up.

Specifically for the medical research field, the Kaplan-Meier (KM) estimator [30] is the most common method used to compute a patient's probability rate of survival. The KM non-parametric method is a statistical model that evaluates the survival trend of patients with a common characteristic, finding the relation between the probability of survival over the observation time in the population analyzed. Equation 1, outlines the process for calculating the KM estimator. This method factors in the occurrence of event d_i and the count of patients n_i who have not yet experienced death or have been censored by time $t_i \leq t$.

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(\frac{n_i - d_i}{n_i} \right) \tag{1}$$

3.2. Differential Privacy

The gold standard definition of privacy widely accepted in the information security community research is provided by the notion of Differential Privacy, introduced in Dwork et al. [9]. A DP mechanism is designed to ensure sensitive data privacy while preserving its utility. In essence, DP adds an appropriately prepared noise level during computation using a so-called privacy budget ε, determining the balance between data privacy and utility. The DP definition is built upon the concept of neighboring datasets, i.e., datasets that can differ at most for only one record. Formally, the definition of ε-DP [9] states that a randomized mechanism $\mathcal{M}$, i.e., a mechanism

that takes an input and outputs a noisy output, is ε-DP if, for any pair of neighboring datasets D and D' and a privacy budget $\varepsilon \in \mathbb{R}^+$, it holds:

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

If a randomized mechanism satisfies ε-DP, then it ensures that the probability of observing any output is almost equal for any neighboring datasets. When facing two similar yet distinct inputs, we expect the output to be the same within a specific probability range, regulated by the privacy budget ε provided. Hence, the mechanism protects users' privacy by ensuring that there is uncertainty about the original data, even if the output is identical for two different inputs. As the definition states, it is possible to infer that the lower ε values are the higher the privacy levels results: if $\varepsilon = 0$, then $\Pr\left[\mathcal{M}(D) \in S\right] = \Pr\left[\mathcal{M}(D') \in S\right]\ \forall S \subset \mathrm{Im}(\mathcal{M})$, i.e., the output of the mechanism does not depend on the input.

An important notion when employing the definition of DP for a randomization mechanism $\mathcal{M}$ is the characterization of its Privacy Loss (PL) measure. Considering two potential inputs neighbor datasets D, D', the PL of the mechanism is defined as the logarithmic ratio between the probabilities of observing the same output O for each input:

$$\mathcal{L}_{\mathcal{M}(D)||\mathcal{M}(D')}(O) = \log\left(\frac{\Pr\left[\mathcal{M}(D) = O\right]}{\Pr\left[\mathcal{M}(D') = O\right]}\right) \tag{2}$$

An important and helpful property for our use-case that links the ε-DP property of a mechanism $\mathcal{M}$ with its measure of PL is provided by Dwork and Roth [31]. Formally, stating that a mechanism $\mathcal{M}$ adheres to ε-DP is equivalent to asserting that the absolute value of PL of the mechanism is upper bounded by ε with probability 1.

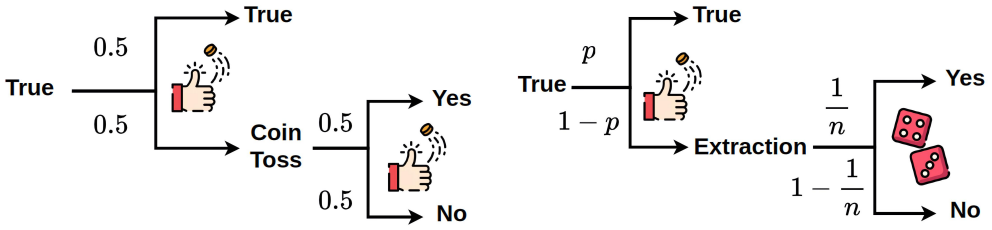
4. Methodology

This Section shows the Revised Randomized Response (RRR) method and its ε-DP property.

4.1. Revised Randomized Response (RRR)

The Randomize Response mechanism was introduced by Warner [13] as a masking technique for protecting the confidentiality of people in survey responses. The Randomize Response mechanism, Figure 1(a), also known as the Direct Encoding mechanism [32], has been used to gather data from survey participants without revealing their real answers. The individuals are asked to respond truthfully or falsely about a property based on a certain condition. This mechanism involves a series of steps, where the individual first flips a fair coin ($p = 0.5$). If the result is "heads", they respond truthfully. If not, they proceed to flip a second fair coin, and if the result of the second coin flip is "heads", the answer is "yes". Otherwise, it returns a "no"answer.

Figure 1(b) illustrates the revised version of the Randomized Response [14, 32]. Unlike the original method, the RRR mechanism adapts inputs that support categories beyond the binary options of "Yes"or "No", extending to adapt multiple categories. The algorithm works in the following manner: given a total of $n > 1$ categories and a coin described by a probability $p \in (0, 1)$ of landing on "heads", the algorithm takes the input category C_i, and selects the



(a) Original Randomize Response mechanism. (b) RRR mechanism.

Figure 1: Schematic comparison between the original Randomize Response with a fair coin ($p = 0.5$) and its Revised version, defined by a coin with bias p, selecting a category C_i among n available ones.

output category through a coin toss process. Similar to the Randomized Response technique, if the coin lands on "heads", the mechanism outputs the true category C_i. Conversely, if the coin lands on "tails", RRR uniformly chooses a category from all available options, varying the Randomize Response mechanism and decreasing the probability of leaking the real category.

4.2. Privacy Properties

Equation 2 measures the privacy loss of $\mathcal{M}$, and the aim is to prove it to be less or equal to ε. As a first step, we evaluate the numerator of that equation. The probability of having the real category as output after receiving category C_i as initial input is provided by the following expression: $\Pr[Resp = C_i | True = C_i] = p + (1-p)\frac{1}{n} = \frac{np+1-p}{n}$. On the other hand, to evaluate the denominator, i.e., the probability of getting C_i as the final response upon collecting $\overline{C_i}$ as input, is estimated by the expression: $\Pr[Resp = C_i | True = \overline{C_i}] = (1-p)(1-\frac{1}{n}) = \frac{n-1-np+p}{n}$.

Therefore, by computing the logarithm of the ratio between the quantities computed above, we formulate the privacy budget ε as a function of p and n, obtaining:

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

The function $\varepsilon(n, p)$ is subject to specific restrictions due to the existential conditions of the logarithmic function. Nonetheless, these limitations can be addressed without significant difficulty. It is crucial to emphasize that the total number of categories should be a positive integer, and the coin's probability parameter p should fall within the real interval $(0, 1)$. Consequentially, by fixing a number of available categories n, it holds the condition in 3, which provides the conclusion for the DP property of the mechanism.

$$\varepsilon > 0 \iff p \in \left(\frac{n-2}{2n-2}, 1\right) \tag{3}$$

In conclusion, we would like to make some important remarks. The condition in 3 states that the privacy budget is regulated by how much the coin is set to be fair, i.e., how much the coin is parameterized to respond truthfully during the obfuscation process. We want to stress that with the same amount of privacy guarantees set by the ε value, the RRR mechanism influences

less the computation of the KM estimator when compared to the LNTE [25] even for small size groups. For instance, the RRR mixes the observations of the events, with the consequence that for different populations, the number of people at risk is more or less the same; thus, the survival curves will not differ too much from one another. On the other hand, the LNTE provides noisy observation as stated by [25]: when the population reaches zero, the LNTE method needs to be truncated. Otherwise, the researcher must rely on a higher privacy budget ε.

5. Results

We discuss the DP approach by simulating a medical research context where we apply the KM estimator, as detailed in Equation 1 to perform SAs. We vary the privacy budget ε to evaluate the privacy-utility trade-off in SA and conduct statistical analysis to confirm our conclusions.

5.1. Experimental Setup

The datasets[1] employed in our experiments represent the typical datasets used in SAs in the medical domain. To ensure a comprehensive evaluation, we used two collections of survival data, commonly used in literature [25, 33] for similar studies, to estimate the SA functions in different privacy settings. We utilized the dataset from Bernard et al. [34], referred to as the IPSS-R dataset, to delineate the group populations under consideration for conducting different SAs. Similarly, the dataset from McGilchrist and Aisbett [35], known as the Kidney dataset, was employed. Such dataset contains survival data about patients who participated in the original studies [34, 35] categorized into different groups based on the diseases or risks.

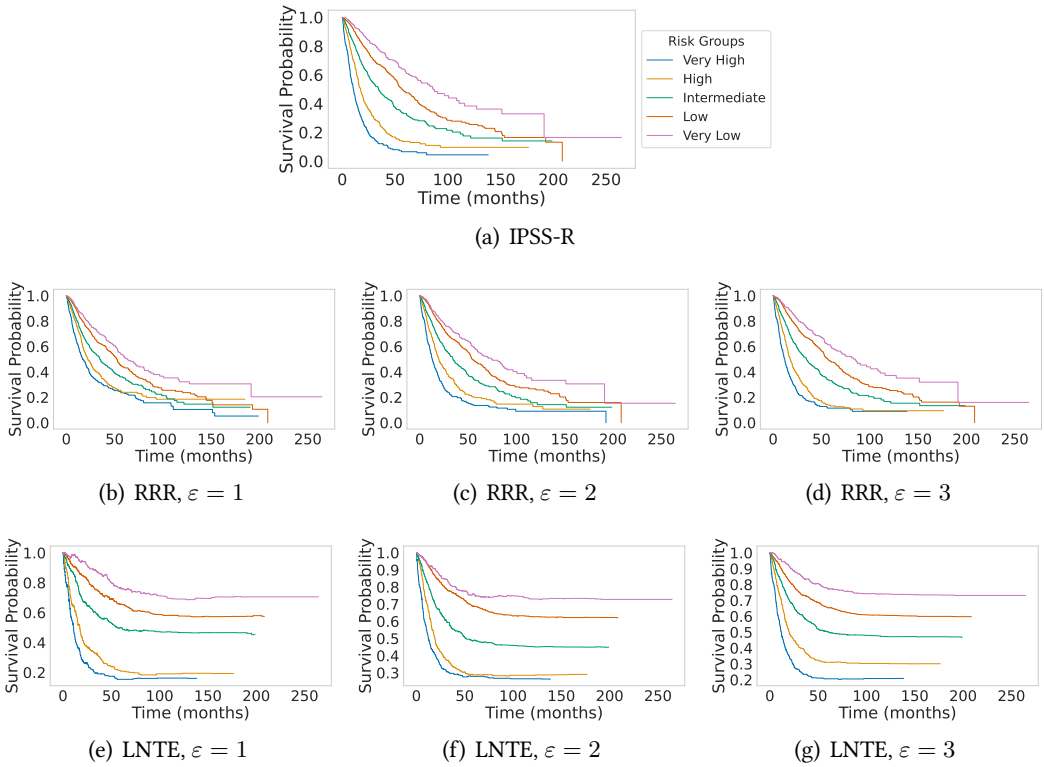
5.2. Survival Curves

Figure 2(a) shows the original KM curves for the IPSS-R dataset[2]. Figures 2(b), 2(c), 2(d) and Figures 2(e), 2(f), 2(g) represents the KM curves obtained using the RRR and LNTE [25] mechanisms, respectively. As done by Gondara and Wang [25], we analyzed the scenarios using three privacy budgets, i.e., $\varepsilon = 1, 2, 3$, to compare how the mechanisms work in the KM estimation.

Our objective was to show how to process sensitive data for computing SAs in a DP manner. Upon applying the RRR and the LNTE mechanisms, we observe a gradual convergence of the survival curves for different values of ε. Specifically, for lower values of the privacy budget, $\varepsilon = 1$, the mechanisms alter the trend of the survival rates across time, making the distinction between risk groups less pronounced and creating a convergence through the "Intermediate" risk population. On the other hand, as ε increases ($\varepsilon = 2, 3$), the survival curves for the RRR method appear to approach those of the original IPSS-R data. However, even at higher ε values, the survival probabilities do not completely align with the original scenario, indicating a residual impact of the privacy-preserving mechanism. Conversely, applying the LNTE, for all ε, indicates a complete change in the survival trends, reducing the utility of the analysis.

[1]The datasets were obtained using the Open Source platform cBioPortal https://www.cbioportal.org/datasets and the dataset available on the R survival package https://cran.r-project.org/web/packages/survival/index.html.

[2]The Kidney Survival Curves, which show similar trends, are omitted due to space limitations but are released along with the code in the GitHub repository at https://github.com/Kekkodf/DP-SurvAnalysis.



(a) IPSS-R

(b) RRR, $\varepsilon = 1$ (c) RRR, $\varepsilon = 2$ (d) RRR, $\varepsilon = 3$

(e) LNTE, $\varepsilon = 1$ (f) LNTE, $\varepsilon = 2$ (g) LNTE, $\varepsilon = 3$

Figure 2: KM Plots in different privacy parametrization settings using data from the IPSS-R dataset.

5.3. Statistical Results

Table 1
Pairwise Log-Rank test statistics using Original and RRR ($\varepsilon = 3$) results from the Kideny dataset, concerning the population based on the kind of disease registered (AN, GN, PKD, Other).

Disease A	**Disease B**	Test Statistic		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

We conducted a Pairwise Log-Rank test on all populations in the Kidney dataset to observe statistical differences in the results obtained. As Table 1 shows, applying the RRR mechanism with $\varepsilon = 3$ shows comparable findings with the originals. In addition, an important insight is

Table 2
Median Survival times with 95% Confidence Interval estimated in the privatized simulations. Where a "-" is placed, it was not possible to compute the Median Survival time.

Dataset	*Mech.*	*Category*	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 3$	*No DP*
IPSS-R	LNTE	V-High	14.73 (9.90, 74.04)	12.07 (9.11, 21.27)	13.02 (10.13,22.88)	10.52
		High	22.26 (15.68, 42.41)	19.66 (15.25, 34.52)	20.52 (16.64, 35.80)	17.29
		Inter.	40.04 (29.10, -)	48.07 (31.59, -)	53.75 (35.80, -)	34.39
		Low	-	-	-	56.68
		V-Low	-	-	-	86.10
	RRR	V-High	15.68 (12.82, 18.87)	14.27 (12.29, 16.27)	11.38 (10.13, 13.05)	10.52
		High	22.32 (19.23, 25.84)	18.38 (16.67, 21.24)	17.65 (15.71, 20.02)	17.29
		Inter.	37.74 (34.19, 43.23)	33.34 (28.57, 38.20)	34.19 (29.10, 38.53)	34.39
		Low	52.93 (48.98, 57.50)	54.87 (51.78, 64.47)	57.47 (52.93, 65.98)	56.68
		V-Low	65.62 (55.53, 82.95)	80.94 (69.76, 100.76)	85.35 (69.76, 108.39)	86.10
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

provided by the p-value fetched: we can see that no p-value is altered so that the null hypothesis of the distribution can be rejected, replicating the same results of the original scenario.

Moreover, we summed the median survival times and their related confidence intervals, Table 2. The analysis reveals that the Kidney dataset's low cardinality has an impact on both methods employed. However, the RRR method returns median times more accurately, which closely resembles the real ones, even in cases where there are high privacy guarantees ($\varepsilon = 1$). On the other hand, the LNTE mechanism disrupts the survival rates of the subject. As a result, calculating median times becomes impractical as the probabilities do not reach the 0.5 threshold.

6. Conclusion

We compared the RRR method with the state-of-the-art for conducting SAs in a DP manner. Our findings suggest that the RRR method is more effective in balancing privacy and utility, maintaining the distribution properties of real results, and ensuring that researchers can still derive important insights from such SAs. Our work contributes to the evolution of privacy-preserving methods in medical research and provides a new comparison for future investigations. As future works, we plan to explore different methods for conducting SAs to gain insights into the privacy-utility trade-offs for such a task. Specifically, we intend to consider Linear and Cox Regression to perform SAs and apply different DP mechanisms to protect patients' privacy.

References

[1] S. Bahri, N. Zoghlami, M. Abed, J. M. R. S. Tavares, Big data for healthcare: A survey, IEEE Access 7 (2019) 7397–7408. doi:10.1109/ACCESS.2018.2889180.

[2] K. M. Batko, A. Slezak, The use of big data analytics in healthcare, J. Big Data 9 (2022) 3. URL: https://doi.org/10.1186/s40537-021-00553-4. doi:10.1186/S40537-021-00553-4.

[3] C. P. Lim, A. Vaidya, Y.-W. Chen, V. Jain, L. C. Jain (Eds.), A Survival Analysis Guide in Oncology, Springer International Publishing, Cham, 2023. URL: https://doi.org/10.1007/978-3-031-11170-9_2. doi:10.1007/978-3-031-11170-9_2.

[4] S. Kuo, M. Ventin, H. Sato, J. M. Harrison, Y. Okuda, M. Qadan, C. R. Ferrone, K. D. Lillemoe, C. F. del Castillo, Common hepatic artery lymph node metastasis in pancreatic ductal adenocarcinoma: An analysis of actual survival, Journal of Gastrointestinal Surgery (2024). doi:https://doi.org/10.1016/j.gassur.2024.02.018.

[5] B. Gudjonsson, E. M. Livstone, H. M. Spiro, Cancer of the pancreas. diagnostic accuracy and survival statistics, Cancer 42 (1978) 2494–2506. doi:https://doi.org/10.1002/1097-0142(197811)42:5<2494::AID-CNCR2820420554>3.0.CO;2-R.

[6] A. Anjum, S. ur Rehman Malik, K.-K. R. Choo, A. Khan, A. Haroon, S. Khan, S. U. Khan, N. Ahmad, B. Raza, An efficient privacy mechanism for electronic health records, Computers & Security 72 (2018) 196–211. URL: https://www.sciencedirect.com/science/article/pii/S0167404817302031. doi:https://doi.org/10.1016/j.cose.2017.09.014.

[7] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, M. Saadi, Big data security and privacy in healthcare: A review, Procedia Computer Science 113 (2017) 73–80. URL: https://www.sciencedirect.com/science/article/pii/S1877050917317015. doi:https://doi.org/10.1016/j.procs.2017.08.292, the 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017) / The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017) / Affiliated Workshops.

[8] A. Almalawi, A. I. Khan, F. Alsolami, Y. B. Abushark, A. S. Alfakeeh, Managing security of healthcare data for a modern healthcare system, Sensors (Basel) 23 (2023) 3612.

[9] C. Dwork, F. McSherry, K. Nissim, A. D. Smith, Calibrating noise to sensitivity in private data analysis, in: S. Halevi, T. Rabin (Eds.), 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*, Springer, 2006, pp. 265–284. URL: https://doi.org/10.1007/11681878_14. doi:10.1007/11681878_14.

[10] W. Liu, Y. Zhang, H. Yang, Q. Meng, A survey on differential privacy for medical data analysis, Annals of Data Science (2023). URL: https://doi.org/10.1007/s40745-023-00475-3. doi:10.1007/s40745-023-00475-3.

[11] K. M. Chong, A. Malip, Bridging unlinkability and data utility: Privacy preserving data publication schemes for healthcare informatics, Computer Communications 191 (2022) 194–207. URL: https://www.sciencedirect.com/science/article/pii/S014036642200144X. doi:https://doi.org/10.1016/j.comcom.2022.04.032.

[12] J. Ficek, W. Wang, H. Chen, G. Dagne, E. Daley, Differential privacy in health research: A scoping review, J Am Med Inform Assoc 28 (2021) 2269–2276.

[13] S. L. Warner, Randomized response: A survey technique for eliminating evasive answer

bias, Journal of the American Statistical Association 60 (1965) 63–69. URL: http://www.jstor.org/stable/2283137.

[14] B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons, D. G. Horvitz, The unrelated question randomized response model: Theoretical framework, Journal of the American Statistical Association 64 (1969) 520–539. URL: http://www.jstor.org/stable/2283636.

[15] T. Ha, T. K. Dang, T. T. Dang, T. A. Truong, M. T. Nguyen, Differential privacy in deep learning: An overview, in: 2019 International Conference on Advanced Computing and Applications (ACOMP), 2019, pp. 97–102. doi:10.1109/ACOMP.2019.00022.

[16] J. Wang, S. Liu, Y. Li, A review of differential privacy in individual data release, International Journal of Distributed Sensor Networks 11 (2015) 259682. URL: https://doi.org/10.1155/2015/259682. doi:10.1155/2015/259682.

[17] X. Zhang, S. Ji, T. Wang, Differentially private releasing via deep generative model, CoRR abs/1801.01594 (2018). URL: http://arxiv.org/abs/1801.01594. arXiv:1801.01594.

[18] H. Bae, D. Jung, H. Choi, S. Yoon, Anomigan: Generative adversarial networks for anonymizing private medical data, in: Pacific Symposium on Biocomputing 2020, Fairmont Orchid, Hawaii, USA, January 3-7, 2020, 2020, pp. 563–574. URL: https://psb.stanford.edu/psb-online/proceedings/psb20/Bae.pdf.

[19] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, C. S. Greene, Privacy-preserving generative deep neural networks support clinical data sharing, Circulation: Cardiovascular Quality and Outcomes 12 (2019) e005122. doi:10.1161/CIRCOUTCOMES.118.005122.

[20] A. Dash, J. Ye, G. Wang, A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines: From medical to remote sensing, IEEE Access 12 (2024) 18330–18357. doi:10.1109/ACCESS.2023.3346273.

[21] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, Y. Pan, Generative adversarial networks: A survey toward private and secure applications, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3459992. doi:10.1145/3459992.

[22] T. T. Nguyên, S. C. Hui, Differentially private regression for discrete-time survival analysis, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 1199–1208. URL: https://doi.org/10.1145/3132847.3132928. doi:10.1145/3132847.3132928.

[23] S. Yu, G. Fung, R. Rosales, S. Krishnan, R. B. Rao, C. Dehing-Oberije, P. Lambin, Privacy-preserving cox regression for survival analysis, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 1034–1042. URL: https://doi.org/10.1145/1401890.1402013. doi:10.1145/1401890.1402013.

[24] K. Chaudhuri, C. Monteleoni, A. D. Sarwate, Differentially private empirical risk minimization, J. Mach. Learn. Res. 12 (2011) 1069–1109. URL: https://dl.acm.org/doi/10.5555/1953048.2021036. doi:10.5555/1953048.2021036.

[25] L. Gondara, K. Wang, Differentially private survival function estimation, in: F. Doshi-Velez, J. Fackler, K. Jung, D. C. Kale, R. Ranganath, B. C. Wallace, J. Wiens (Eds.), 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*, PMLR,

2020, pp. 271–291. URL: http://proceedings.mlr.press/v126/gondara20a.html.

[26] T. G. Clark, M. J. Bradburn, S. B. Love, D. G. Altman, Survival analysis part i: Basic concepts and first analyses, British Journal of Cancer 89 (2003) 232–238. URL: https://doi.org/10.1038/sj.bjc.6601118. doi:10.1038/sj.bjc.6601118.

[27] L. L. Johnson, Chapter 26 - an introduction to survival analysis, in: J. I. Gallin, F. P. Ognibene, L. L. Johnson (Eds.), Principles and Practice of Clinical Research (Fourth Edition), fourth edition ed., Academic Press, Boston, 2018, pp. 373–381. URL: https://www.sciencedirect.com/science/article/pii/B9780128499054000265. doi:https://doi.org/10.1016/B978-0-12-849905-4.00026-5.

[28] B. Bieszk-Stolorz, Application of the survival analysis methods in contemporary economics on the example of unemployment, in: K. Nermend, M. Łatuszyńska (Eds.), Experimental and Quantitative Methods in Contemporary Economics, Springer International Publishing, Cham, 2020, pp. 115–131.

[29] N. R. Latimer, Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data, Med. Decis. Making 33 (2013) 743–754.

[30] E. L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, Journal of the American Statistical Association 53 (1958) 457–481. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452. doi:10.1080/01621459.1958.10501452.

[31] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Foundations and Trends® in Theoretical Computer Science 9 (2014) 211–407. URL: http://dx.doi.org/10.1561/0400000042. doi:10.1561/0400000042.

[32] T. Wang, J. Blocki, N. Li, S. Jha, Locally differentially private protocols for frequency estimation, in: E. Kirda, T. Ristenpart (Eds.), 26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017, USENIX Association, 2017, pp. 729–745. URL: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/wang-tianhao.

[33] P.-C. Bürkner, brms: An R Package for Bayesian Multilevel Models Using Stan, Journal of Statistical Software 80 (2017) 1–28. URL: https://www.jstatsoft.org/index.php/jss/article/view/v080i01. doi:10.18637/jss.v080.i01.

[34] 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, E. Papaemmanuil, Molecular international prognostic scoring system for myelodysplastic syndromes, NEJM Evidence 1 (2022) EVIDoa2200008. URL: https://evidence.nejm.org/doi/abs/10.1056/EVIDoa2200008. doi:10.1056/EVIDoa2200008.

[35] C. A. McGilchrist, C. W. Aisbett, Regression with frailty in survival analysis, Biometrics 47 (1991) 461–466. URL: http://www.jstor.org/stable/2532138.