Differentially Private Survival Function Estimation

Anonymous authors
Paper under double-blind review

Abstract

Survival function estimation is used in many disciplines, but it is most common in medical analytics in the form of the Kaplan-Meier estimator. Sensitive data (patient records) is used in the estimation without any explicit control on the information leakage, which is a significant privacy concern. We propose a first differentially private estimator of the survival function and show that it can be easily extended to provide differentially private confidence intervals and test statistics without spending any extra privacy budget. We further provide extensions for differentially private estimation of the competing risk cumulative incidence function. Using nine real-life clinical datasets, we provide empirical evidence that our proposed method provides good utility while simultaneously providing strong privacy guarantees.

1 Introduction

A patient progresses from HIV infection to AIDS after 4.5 years. A study using the patient's data publishes the survival function estimates (a standard practice in clinical research). An adversary, with only access to the published estimates (even in the form of survival function plots), can reconstruct user-level data (Wei & Royston, 2018; Fredrikson et al., 2014). Effectively leading to the disclosure of sensitive information. This is just one scenario. The survival function is used for modeling any time to an event, taking into account that some subjects will not experience the event at the time of data collection. The survival function is used in many domains, some examples are the duration of unemployment (in economics); time until the failure of a machine part (in engineering); time to disease recurrence, time to infection, time to death (in healthcare); etc.

Our personal healthcare information is the most sensitive private attribute, protected by law, violations of which carry severe penalties. And as the initial example suggests, of all application areas, information leakage in the healthcare domain is the most serious issue and is our focus in this study. For estimation of the survival function, we focus on the Kaplan-Meier's (KM) (Kaplan & Meier, 1958) non-parametric method. KM's method is ubiquitous in clinical research. A quick search of the term on PubMed[1] yields 109,421 results. It is not an overstatement to say that almost every clinical study uses KM's method to report summary statistics on their cohort's survival. Statistical agencies around the world use this method to report on the survival of the general population or specific disease-related survival estimates.

To best of our knowledge, there does not exist any model that can provide formal privacy guarantees for estimation of survival function using the KM method. The only related work is by Nguyên & Hui (2017), which uses the output and objective perturbation for regression modeling of discrete time to event data. The approach is limited to "multivariate" regression models and cannot be directly used to estimate survival function in a differentially private fashion. One can argue that generative models such as the differentially private generative adversarial networks (Xie et al., 2018; Zhang et al., 2018; Triastcyn & Faltings, 2018; Beaulieu-Jones et al., 2017; Yoon et al., 2019) can be trained to generate differentially private synthetic data. Which can then be used to estimate the survival function. But, GANs do not generalize well to the datasets typically encountered for our use-case (very small sample size (can be less than a hundred), highly constrained dimensionality ($d \in [2, 3]$), a mixture of categorical and continuous variables, no data pre-processing allowed, etc.).

[1] A free search engine indexing manuscripts and abstracts for life sciences and other biomedical topics. Link - `https://www.ncbi.nlm.nih.gov/pubmed/`

We propose the first differentially private method for estimating the survival function based on the KM method. Grounded by the core principles of differential privacy, our method guarantees the differentially private estimation of the survival function. Also, we show that our method easily extends to provide differentially private confidence intervals and differentially private test statistics (for comparison of survival function between multiple groups) without any extra privacy cost. We further extend our method for differentially private estimation of the competing risk cumulative incidence function (another popular estimate in clinical research). Using nine real-life clinical datasets, we provide empirical evidence that our proposed method provides good utility while simultaneously providing strong privacy guarantees. Lastly, we release our method as an R[2] (R Core Team, 2018) package for rapid accessibility and adoption.

2 Preliminaries and Technical Background

We use this section to introduce the concepts central to the understanding of our method.

2.1 Survival Function

The survival function is used to model time to event data, where the event may not have yet occurred (but the probability of occurrence is non-zero). Such as for HIV infection to AIDS timeline data, at the end of the follow-up period, some patients would have progressed (our event of interest), while others would not have yet progressed (censored observations). Accounting for censored observations (patients that never experience the event during our follow-up) is the central component in the estimation of the survival function. Formally,

$$S(t) = P(T > t) = \int_t^\infty f(u)\, du = 1 - F(t) \tag{1}$$

this gives the probability of not having an event just before time t, or more generally, the probability that the event of interest has not occurred by time t.

In practice, survival function can be estimated using more than one approach. Several parametric methods (that make assumptions on the distribution of survival times) such as the ones based on the exponential, Weibull, Gompertz, and log-normal distributions are available. Or one can opt for the most famous and most often used non-parametric method (Kaplan-Meier's method (Kaplan & Meier, 1958)), which does not assume how the probability of an event changes over time. Our focus in this paper is the latter, which has become synonymous with survival models in clinical literature. KM estimator of the survival function is defined as follows

$$\hat{S}(t) = \prod_{j:t_j \leq t} \frac{r_j - d_j}{r_j} \tag{2}$$

where $t_j, (j \in 1, \cdots, k)$ is the set of k distinct failure times (not censored), d_j is the number of failures at t_j, and r_j are the number of individuals "at risk" before the j-th failure time. We can see that the function $\hat{S}(t)$ only changes at each failure time, not for censored observations, resulting in a "step" function (the characteristic feature of KM estimate).

2.2 Differential Privacy

Differential privacy (Dwork et al., 2006) provides provable privacy notion, with the intuition that a randomized algorithm behaves similarly on similar input datasets. Formally,

Definition 1. (Differential privacy (Dwork et al., 2006)) *A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ preserves (ϵ, δ)-differentially privacy if for all $\mathcal{S} \subseteq$ Range$(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $||x - y||_1 \leq 1$, we have*

$$Pr[\mathcal{M}(x) \in \mathcal{S}] \leq e^\epsilon Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta \tag{3}$$

where the two datasets (x, y) only differ in any one row (neighboring datasets) and the probability space is over the coin flips of the mechanism $\mathcal{M}$. If $\delta = 0$, we have "pure ϵ-differential privacy". Smaller (ϵ, δ) provide stronger privacy guarantees.

[2] Most often used programming language in medical statistics

Differential privacy has some interesting properties. We briefly introduce the main property that is crucial to our proposal of differentially private survival function estimation. That is, the *post-processing*, formally

Theorem 1. (Post processing (Dwork et al., 2006)) *Let $\mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to R$ be a randomized algorithm that is (ϵ, δ)-differentially private. Let $f : R \to R'$ be an arbitrary randomized mapping. Then $f \circ \mathcal{M} : \mathbb{N}^{|\mathcal{X}|} \to R'$ is (ϵ, δ)-differentially private.*

Theorem 1 states that differential privacy is immune to post-processing. That is, an adversary acting only on the output of $\mathcal{M}$, cannot increase the privacy loss. This notion is central to our approach and we will revisit it in the following sections.

3 Differentially Private Estimation of Survival Function

Now we introduce our method for differentially private estimation of the survival function using the Kaplan-Meier's method. We follow the basic principles of differential privacy to ensure that our estimate of the survival function is differentially private. We subsequently show that following our simple approach, it is possible to estimate a wide variety of accompanying statistics (such as the confidence intervals, comparison test statistics, etc.) in a differentially private way without spending any extra privacy budget.

3.1 Estimation

Before we begin, we recap some of the notations introduced in Section 2.1. We have a vector of time points ($t_j, j \in \{1, \cdots, k\}$), and for each time point, we have a corresponding number of subjects at risk r_j (number of subjects not experiencing a progression up to that time point), and we have the number of subjects experiencing the event at that time point (number of progressions), which we denote as d_j.

We first create a dataset (a matrix) where each row has the data on the number of events (d_j) and the number at risk (r_j) for each unique time point (t_j). Let's denote this matrix by M. Then using the L_1 sensitivity ($\mathcal{S}$) of M, we draw a noise matrix Z from the Laplace distribution ($Lap(\mathcal{S}/\epsilon)$), where ϵ is the privacy parameter and Z is of the same size as M. We then create a differentially private version of M by adding Z, that is, $M' = M + Z$. All subsequent calculations use M'. We succinctly present our method as Algorithm 1.

Algorithm 1 Differentially Private Estimation of $\hat{S}(t)$

```
procedure DP($\hat{S}(t)$)
    Create a matrix $M$; $[r_j, d_j] \in M$; for every $t_j$
    $M' = M + Lap(\mathcal{S}/\epsilon)$; $[r'_j, d'_j] \in M'$
    $\hat{S}'(t) = \prod_{j:t_j \leq t} \frac{r'_j - d'_j}{r'_j}$
    return $\hat{S}'(t)$
end procedure
```

We use this paragraph to briefly discuss Algorithm 1. We begin with the noticeable simplicity of the procedure, that is, the minimal changes required to the original estimation procedure to make it differentially private. This further boosts the accessibility of our differentially private version (it can be implemented using any readily available software package). Also, the required changes for differential privacy come with no computational overhead compared to the original estimation (our method is computationally cheap). Below we provide the formal privacy guarantees and further details on how this method can be easily extended for differentially private estimation of "other" associated statistics.

3.2 Privacy Guarantees

Now we are ready to formally state the differential privacy guarantees of our proposed method. Before we state our main theorem, we start with a supporting Lemma for establishing the global L_1 sensitivity ($\mathcal{S}$) of our method.

Lemma 1. *L_1 sensitivity ($\mathcal{S}$) of M is two.*

Proof. As M only contains count variables for the number of events and number at risk for each unique time point. Adding or removing any single individual can change the counts by at most two (that is being in at-risk group and having an event). □

Theorem 2. *Algorithm 1 is ϵ-differentially private.*

Proof. Sketch: We have established the L_1 sensitivity of M. Using it to add Laplace noise ($M' = M + Lap(2/\epsilon)$) makes sure M' is differentially private and so are its individual components (that are r'_j, d'_j). Using (r'_j, d'_j) to calculate the survival function (Eqn. 2) ensures that the estimated function is differentially private by the post-processing theorem (Dwork & Roth, 2014). Complete formal proof is provided in the Appendix. □

4 Extending to Other Estimates

As mentioned in the introduction, one of the advantages of our approach is its easy extension to other essential statistics often required and reported along with the estimates of the survival function. Such as the confidence intervals, test statistics for comparing the survival function distributions, etc. Here we formally define the extensions with their privacy guarantees.

4.1 Confidence Intervals and Test Statistics

When reporting survival function estimates, it is often required to include the related confidence intervals, reported to reflect the uncertainty of the estimate. And for group comparison, such as comparing the infection rates between two treatment arms of a clinical trial, hypothesis testing is used with the help of test statistic. So, it is of paramount interest to provide the differentially private counterparts of both (confidence intervals and test statistics). We start with the confidence intervals.

Confidence Intervals for survival function estimates are calculated in a "point-wise" fashion, that is, they are calculated at discrete time-points whenever an event is observed (for the same time points at which the survival function changes its value). We start with proving that the calculations required for obtaining confidence intervals are differentially private following the changes made to the data in Algorithm 1.

Theorem 3. *Confidence Intervals for $\hat{S}'(t)$ are ϵ-differentially private.*

Proof. There are more than one type of confidence intervals available for the survival function. Here we focus on the most often used Greenwood's (Greenwood et al., 1926) linear-point-wise confidence intervals.

Greenwood's formula for the confidence intervals is given as

$$\hat{S}(t) \pm z_{1-\alpha/2}\sigma_S(t) \tag{4}$$

where

$$\sigma_s^2(t) = \hat{V}[\hat{S}(t)] \tag{5}$$

and

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)} \tag{6}$$

Replacing by their respective differentially private counterparts from Algorithm 1.

$$\hat{V}'[\hat{S}(t)] = \hat{S}'(t)^2 \sum_{t_j \leq t} \frac{d'_j}{r'_j(r'_j - d'_j)} \tag{7}$$

estimate for $\hat{V}'[\hat{S}(t)]$ is now differentially private, using it in conjunction with $\hat{S}'(t)$ makes the confidence intervals differentially private by the post-processing theorem (Dwork et al., 2006). □

As we don't need any additional access to the sensitive data for calculating confidence intervals. Hence, calculating and providing differentially private confidence intervals with the differentially private survival function estimates does not incur any additional privacy cost. In other words, we get the differentially private confidence intervals for free.

Hypothesis tests are often used to compare the distribution of survival function estimates between groups. For example: To compare infection rates between two treatment arms of a clinical trial. Most often used statistical test in such scenarios is the Logrank test (Mantel, 1966). Below we show that using our method (Algorithm 6), the hypothesis testing using the Logrank test is differentially private.

Theorem 4. *Hypothesis test for $\hat{S}'(t)$ is ϵ-differentially private.*

Proof. Logrank test statistic (Z) is given as

$$Z = \frac{\sum_{j=1}^{k}(O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{k} V_j}} \tag{8}$$

where O_{1j} are observed number of failures (events) (d_{1j}) and E_{1j} are the expected number of failures at time j in group 1, we have

$$E_{1j} = d_j \frac{r_{1j}}{r_j} \tag{9}$$

and V_j is the variance, given as

$$V_j = \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)} \tag{10}$$

Replacing the corresponding quantities by their differentially private counterparts using Algorithm 1, we get

$$V'_j = \frac{r'_{1j} r'_{2j} d'_j (r'_j - d'_j)}{r'^2_j (r'_j - 1)} \tag{11}$$

which makes V'_j differentially private as no other sensitive information is required for its estimation.

Using it in conjunction with O_{1j} and E_{1j}, which can be made differentially private following the same argument, makes the test statistic Z differentially private by the post-processing theorem (Dwork et al., 2006). □

The calculation, again being the case of standard post-processing on differentially private data does not add to our overall privacy budget. Hence, after using Algorithm 1, we can output the related confidence intervals and the test statistic without spending any additional privacy budget.

4.2 Extending to Competing Risks (Cumulative Incidence)

In certain scenarios, we can have more than one type of event. Using our prior example of HIV infection, we might have a scenario where patients die before progression to AIDS, making the observation of progression impossible. Such events (death) that preclude any possibility of our event of interest (progression) are known as competing events. Competing events are a frequent occurrence in clinical data and require specialized estimates that take this phenomenon into account, without which our estimates will be biased. One such estimate is the competing risk cumulative incidence, which is also the most widely used and reported estimate in the literature, akin to the KM estimate, but for competing events.

Here we show that using Algorithm 1, we can easily extend differential privacy to competing risk scenarios.

Theorem 5. *Competing risk cumulative incidence using our method is ϵ-differentially private.*

Proof. Cumulative incidence extends Kaplan-Meier estimator and is given by

$$\hat{I}_j(t) = \sum_{i:t_i<t} \hat{S}(t_i)\frac{d_{ij}}{n_i} \tag{12}$$

where d_{ij} is the number of events of type j at time $t_{(i)}$ and $\hat{S}(t_i)$ is the standard Kaplan-Meier estimator at time $t_{(i)}$.

Replacing associated quantities with their differentially private counterparts (using same reasoning as Algorithm 1).

$$\hat{I}_j(t)' = \sum_{i:t_i<t} \hat{S}(t_i)'\frac{d'_{ij}}{n'_i} \tag{13}$$

Its not hard to see that $\hat{I}_j(t)'$ is differentially private by the post-processing theorem. □

4.3 More Extensions

Further statistics associated with the cumulative incidence such as the confidence intervals and hypothesis tests, hazard function and hazard ratios, etc. that directly depend on the quantities made differentially private using Algorithm 1 can be similarly argued to be differentially private. Another popular extension that we easily get from our method is the differentially private version of the Nelson-Aalen estimate of the cumulative hazard (Nelson, 1972; 1969; Aalen, 1978). Which is simply $\hat{H}'(t) = \sum_{t_j \leq t} d'_j/r'_j$, or can be estimated directly from its relationship with the survival function ($\hat{S}'(t) = \exp(-\hat{H}'(t))$).

5 Empirical Evaluation

Here we present the empirical evaluation of our method on nine real-life clinical datasets of varying properties. We start with the dataset description.

5.1 Datasets

Nine real-life clinical datasets with time to event information are used to evaluate our proposed method. Dataset details are provided in Table 1. For space constraints, we provide further details (dataset properties, pre-processing, group comparison details for hypothesis tests, etc.) in the Appendix.

Table 1: Datasets used for evaluation of our proposed method, observations are the number of observations (rows) in the dataset. Wide variety of datasets are used to simulate real-world clinical scenarios.

Dataset	Observations
Cancer	228
Gehan	42
Kidney	76
Leukemia	23
Mgus	1384
Myeloid	646
Ovarian	26
Stanford	184
Veteran	137

5.2 Setup and Comparison

Privacy budget (ϵ) is varied as reported in the results. Being a "non-trainable" model, there are no train/test splits and results are reported on the complete dataset as an average of ten runs. All experiments are performed in R (R Core Team, 2018) with the source code and the datasets made publicly available on GitHub and as an R package[3].

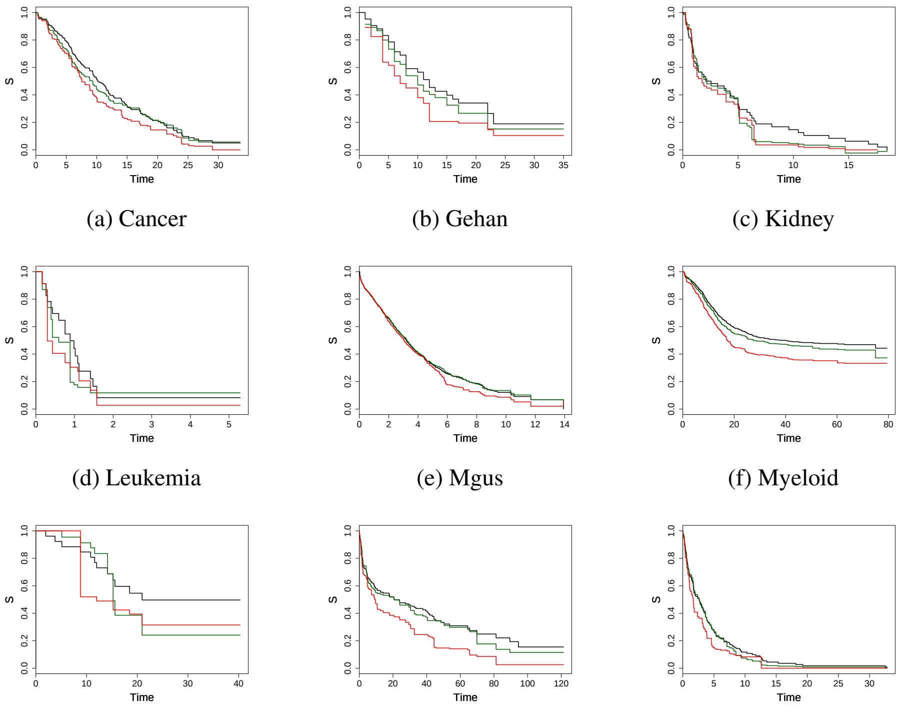
As there is no current method for producing differentially private estimates of the survival function. We compare our approach to the original "non-private" version. This provides us with a comparison to the upper bound (we cannot get better than the non-noisy version). Good utility in comparison with the original non-perturbed version will add credibility to our claim of high utility and will encourage practitioners to adopt our method for practical use.

5.3 Results

Now we present the outcome of our evaluation on nine real-life datasets. We start with the estimation of the differentially private survival function and then move on to the evaluation of the extensions (confidence intervals, test statistic, etc.).

5.3.1 Main Results

For the differentially private estimation of the survival function (our primary goal), Figure 1 shows the results. We can see that our privacy-preserving estimation (green line) faithfully estimates the survival function (black line), with little utility loss. As expected, estimation deteriorates with increased privacy budget (orange line).



(a) Cancer

(b) Gehan

(c) Kidney

(d) Leukemia

(e) Mgus

(f) Myeloid

(g) Ovarian

(h) Stanford

(i) Veteran

Figure 1: Differentially private estimation of the survival function: Followup time is on the X-axis and the probability of survival is on the Y-axis. The black line is the original function estimate, the green line is the differentially private estimate with $\epsilon = 2$, and the orange line is the differentially private estimate with $\epsilon = 1$. We observe that our method provides good utility while protecting an individual's privacy. Small sample sized datasets fare worse compared to larger datasets.

[3] Public link omitted to respect anonymous review, available as an anonymous file on `https://bit.ly/2kNHC1J`

An observation worth making is that as the dataset size gets smaller (a small number of events; as in ovarian, Leukemia, Gehan), the utility of our differentially private estimation gets worse. Which is intuitive from the differential privacy point of view, because to protect an individual's privacy in a small dataset, we will need to add large noise (large perturbation). Whereas for moderate to medium-sized datasets, our differentially private estimation provides good results, even for the high privacy regime.

5.3.2 Median Survival, Confidence Intervals, Test Statistic, and Cumulative Incidence

An important estimate often reported with survival function is the median survival time and its associated confidence intervals. Median survival time is defined as the time point when the survival function attains the value of 0.5, confidence intervals for the survival function at that time point serve as the confidence intervals of the median survival. Table 2 shows the results. For "Median Survival (95% CI)", we see that our method estimates the median with high precision, even for high privacy regime. For some cases due to high survival (as is the case with Myeloid and Ovarian datasets), it is not possible to estimate the upper bounds on the confidence intervals, that is why they are marked as "NA". We see a similar trend as we saw with results in Figure 1, our precision increases with increasing dataset size, an acceptable trade-off for individual-level privacy protection.

Table 2: Median Survival with associated confidence intervals and the test statistic for comparing two survival distributions. ϵ is the privacy budget for our method and "No privacy" are the results from the non-noisy model. Our method provides "close" estimates to the original non-noisy values.

	Median Survival(95% CI)			Test Statistic (Z)		
Dataset	$\epsilon = 2$	$\epsilon = 1$	No Privacy	$\epsilon = 2$	$\epsilon = 1$	No Privacy
Cancer	9.3 (7.4, 10.2)	7.6 (7.0, 8.9)	10.2 (9.4, 11.9)	12.8	13.6	11.4
Gehan	10.0 (6.0, 15.0)	7.0 (4.0, 11.0)	12.0 (8.0, 22.0)	17.1	27.4	16.3
Kidney	2.1 (1.3, 4.3)	1.7 (1.0, 3.9)	2.6 (1.3, 5.0)	12.1	27.8	7.0
Leukemia	0.6 (0.2, 0.9)	0.3 (0.1, 0.9)	0.9 (0.6, 1.5)	4.1	4.4	3.6
Mgus	3.2 (3.0, 3.4)	3.0 (2.9, 3.3)	3.2 (3.0, 3.4)	5.9	4.8	7.6
Myeloid	27.5 (22.9, NA)	17.3 (16.3, NA)	40.1 (27.1, NA)	11.3	15.6	8.7
Ovarian	15.2 (14.2, NA)	11.9 (8.8, NA)	21.0 (15.2, NA)	2.1	6.5	1.1
Stanford	20.7 (8.6, 30.5)	9.2 (6.3, 10.8)	20.7 (10.8, 40.5)	5.9	7.7	5.6
Veteran	2.6 (1.8, 3.4)	1.7 (1.1, 1.8)	2.6 (1.7, 3.5)	1.0	3.4	0.02

For the test statistic[4], in Table 2, we observe that our differentially private estimation performs at par with the original "non-noisy" estimation, even for the high privacy regime. The test statistic (Z) follows the χ^2 distribution with one degree of freedom. Using it to derive the p-values, we observe that none of the differentially private estimates change statistical significance threshold (at 0.05 level). That is, none of the differentially private estimates make the "non-noisy" statistically significant results non-significant or vice-versa.

For cumulative incidence, we use two new datasets with competing risk information. Results are similar with the estimation of competing risk cumulative incidence, that is, our proposed method provides good utility while protecting an individual's privacy. Our method provides faithful estimation even at high privacy regime. For space constraints, detailed results are presented in the Appendix.

6 Related Work

Much work has been done in the intersection of statistical modeling and differential privacy, including many works proposing different differentially private methods for regression modeling (Sheffet, 2017; Jain et al., 2012; Zhang et al., 2012; Yu et al., 2014; Chaudhuri et al., 2011). Using the same

[4] Obtained from comparing the survival distribution of different groups in the dataset, group details are provided in the Appendix.

principles, Nguyên & Hui (2017) further developed a differentially private regression model for survival analysis. This approach is limited to the "multivariate" regression models and cannot be used for direct differentially private estimation of the survival function. Differentially private generative models such as the differentially private generative adversarial networks (Xie et al., 2018; Zhang et al., 2018; Triastcyn & Faltings, 2018; Beaulieu-Jones et al., 2017; Yoon et al., 2019) have been recently proposed. But, as discussed in the introduction, they are not suitable for generating data for survival function estimation.

7 Conclusion

We have presented the first method for differentially private estimation of the survival function and we have shown that our proposed method can be easily extended to differentially private estimation of "other" often used statistics such as the associated confidence intervals, test statistics, and the competing risk cumulative incidence. With extensive empirical evaluation on nine real-life datasets, we have shown that our proposed method provides good privacy-utility trade-off.

References

Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pp. 701–726, 1978.

Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *BioRxiv*, pp. 159756, 2017.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

David Roxbee Cox. *Analysis of survival data*. Routledge, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3&#8211;4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pp. 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-32731-2, 978-3-540-32731-8. doi: 10.1007/11681878_14. URL http://dx.doi.org/10.1007/11681878_14.

John H Edmonson, Thomas Richard Fleming, DG Decker, GD Malkasian, EO Jorgensen, JA Jefferies, MJ Webb, and LK Kvols. Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. *Cancer treatment reports*, 63(2):241–247, 1979.

Luis A Escobar and William Q Meeker Jr. Assessing influence in regression analysis with censored data. *Biometrics*, pp. 507–528, 1992.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pp. 17–32, 2014.

Major Greenwood et al. A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*, (33), 1926.

Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *Conference on Learning Theory*, pp. 24–1, 2012.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

W Ray Kim, Terry M Therneau, Joanne T Benson, Walter K Kremers, Charles B Rosen, Gregory J Gores, and E Rolland Dickson. Deaths on the liver transplant waiting list: an analysis of competing risks. *Hepatology*, 43(2):345–351, 2006.

Robert A Kyle. benign monoclonal gammopathyafter 20 to 35 years of follow-up. In *Mayo Clinic Proceedings*, volume 68, pp. 26–36. Elsevier, 1993.

Charles Lawrence Loprinzi, John A Laurie, H Sam Wieand, James E Krook, Paul J Novotny, John W Kugler, Joan Bartel, Marlys Law, Marilyn Bateman, and Nancy E Klatt. Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group. *Journal of Clinical Oncology*, 12(3):601–607, 1994.

Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50:163–170, 1966.

CA McGilchrist and CW Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47(2): 461–466, 1991.

Rupert G Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

Wayne Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1): 27–52, 1969.

Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

Thông T Nguyên and Siu Cheung Hui. Differentially private regression for discrete-time survival analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1199–1208. ACM, 2017.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

Or Sheffet. Differentially private ordinary least squares. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3105–3114. JMLR. org, 2017.

Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2013.

Aleksei Triastcyn and Boi Faltings. Generating differentially private datasets using gans. *arXiv preprint arXiv:1803.03148*, 2018.

Yinghui Wei and Patrick Royston. Reconstructing time-to-event data from published kaplan–meier curves. *The Stata Journal*, 17(4):786–802, 2018.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.

Fei Yu, Michal Rybar, Caroline Uhler, and Stephen E Fienberg. Differentially-private logistic regression for detecting multiple-snp association in gwas databases. In *International Conference on Privacy in Statistical Databases*, pp. 170–184. Springer, 2014.

Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment*, 5(11):1364–1375, 2012.

Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model. *arXiv preprint arXiv:1801.01594*, 2018.

A Appendix

A.1 Dataset Details

Here we provide details on the datasets used for evaluation.

1. Cancer: It pertains to the data on survival in patients with advanced lung cancer from the North Central Cancer Treatment Group Loprinzi et al. (1994). Survival time in days is converted into months. Groups compared are males and females.
2. Gehan: This is the dataset from a trial of 42 leukemia patients Cox (2018). Groups compared are the control and treatment groups.
3. Kidney: This dataset is on the recurrence times to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment McGilchrist & Aisbett (1991). Time is converted into months and groups compared are males and females.
4. Leukemia: This pertains to survival in patients with Acute Myelogenous Leukemia Miller Jr (2011). Time is converted into months and groups compared are whether maintenance chemotherapy was given or not.
5. Mgus: This dataset is about natural history of subjects with monoclonal gammopathy of undetermined significance (MGUS) Kyle (1993). Time is converted into months and groups compared are males and females.
6. Myeloid: Dataset is based on a trial in acute myeloid leukemia. Time is converted into months and groups compared are the two treatment arms.
7. Ovarian: This dataset pertains to survival in a randomized trial comparing two treatments for ovarian cancer Edmonson et al. (1979). Time is converted into months and groups compared are the different treatment groups.
8. Stanford: This dataset contains the Stanford Heart Transplant data Escobar & Meeker Jr (1992). Time is converted into months and groups compared are the age groups (above and below median).
9. Veteran: This dataset has information from randomized trial of two treatment regimens for lung cancer Kalbfleisch & Prentice (2011). Time is converted into months and groups compared are the treatment arms.

A.2 Extending to Competing Risk (Cumulative Incidence)

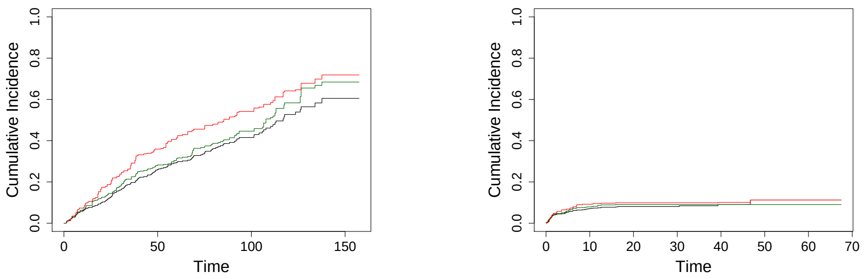
For empirical evaluation in a competing risk scenario, we use two datasets that have more than one type of event. First is from a clinical trial for primary biliary cirrhosis (PBC) of the liver (Therneau & Grambsch, 2013). With the event variable being receipt of a liver transplant, censor, or death; our event of interest is the transplant, and death here is a competing event. The second dataset has the data on the subjects on a liver transplant waiting list from 1990-1999, and their disposition: received a transplant (event of interest), died while waiting (competing risk), or censored (Kim et al., 2006).

Figure 2 shows the results (cumulative incidence is the opposite of survival function, so the plots go upward). We observe that our differentially private extension does an excellent job of differentially private estimation of the competing risk cumulative incidence function while providing strong privacy guarantees.

A.3 Proofs

Theorem 6. *Algorithm 1 is ϵ-differentially private.*

Proof. Let $M \in \mathbb{R}^d$ and $M^* \in \mathbb{R}^d$, such that the L_1 sensitivity, $\mathcal{S}$, is $||M - M^*||_1 \leq 1$, and let $f(.)$ denote some function, $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$. Let p_M denote the probability density function of $\mathcal{Z}(M, f, \epsilon)$, and let p_{M^*} denote the probability density function of $\mathcal{Z}(M^*, f, \epsilon)$, we compare both at some arbitrary point $q \in \mathbb{R}^k$.



(a) PBC

(b) Transplant

Figure 2: Extending differentially private estimation to competing risk cumulative incidence (cumulative incidence is the opposite of survival function, so the plots go upward). Black is the original, unperturbed estimate. Green is with $\epsilon = 2$ and orange is with $\epsilon = 1$. We can see that our method does a good job of estimating competing risk cumulative incidence while providing strong privacy guarantees.

$$\begin{aligned}
\frac{p_M(q)}{p_{M^*}(q)} &= \prod_{i=1}^{k} \left(\frac{\exp(-\frac{\epsilon |f(M)_i - q_i|}{\Delta f})}{\exp(-\frac{\epsilon |f(M^*)_i - q_i|}{\Delta f})} \right) \\
&= \prod_{i=1}^{k} \exp \left(\frac{\epsilon(|f(M^*)_i - q_i| - |f(M)_i - q_i|)}{\Delta f} \right) \\
&\leq \prod_{i=1}^{k} \exp \left(\frac{\epsilon |f(M)_i - f(M^*)_i|}{\Delta f} \right) \\
&= \exp \left(\frac{\epsilon ||f(M) - f(M^*)||_1}{\Delta f} \right) \\
&\leq \exp(\epsilon)
\end{aligned} \tag{14}$$

last inequality follows from the definition of sensitivity $\mathcal{S}$

As our function estimation uses everything from M' (our differentially private version of M) and nothing else from the sensitive data, our survival function estimation is differentially private by the post-processing Theorem (Dwork & Roth, 2014). □