Differentially Private Survival Function Estimation

Lovedeep Gondara LGONDARA@SFU.CA
Department of Computing Science
Simon Fraser University
Burnaby, BC, Canada

Ke Wang WANGK@CS.SFU.CA
Department of Computing Science
Simon Fraser University
Burnaby, BC, Canada

Editor: Editor's name

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, Nelson-Aalen's estimator for the hazard function, etc. Using eleven 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 and Royston, 2018; Fredrikson et al., 2014), effectively leading to the disclosure of sensitive information (Dinur and Nissim, 2003). 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 the next purchase (for churn identification in business); 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 and Meier, 1958) non-parametric method. KM's method is

ubiquitous in clinical research. A quick search of the term on PubMed[1] yields more than 110,000 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 method that can provide formal privacy guarantees for estimation of survival function using the KM method. The only related work is by Nguyên and Hui (2017), which uses the output and objective perturbation for regression modeling of discrete time-to-event data. However, this approach is limited to discrete-time models, whereas our focus is on the continuous-time paradigm. Furthermore, this approach is limited to "multivariate" regression models; and hence 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 and Faltings, 2018; Beaulieu-Jones et al., 2017; Esteban 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 (scaling, etc.) allowed, etc.).

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 and the hazard function using the Nelson-Aalen estimator (Nelson, 1972, 1969; Aalen, 1978) (other popular estimates in clinical research). Using eleven 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

This work assumes familiarity with survival analysis and differential privacy. We present some introductory concepts below, and direct readers to Kleinbaum and Klein (2010) and Kalbfleisch and Prentice (2011) for a detailed exposition on survival analysis and Dwork and Roth (2014) for a detailed introduction to differential privacy.

2.1. Survival Function

The survival function is used to model time to event/time to failure data (we use event and failure interchangeably throughout the paper), where the event may not have yet occurred

1. A free search engine indexing manuscripts and abstracts for life sciences and other biomedical topics. Link - https://www.ncbi.nlm.nih.gov/pubmed/
2. Most often used programming language in medical statistics

(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 (an event/our event of interest), while others would not have yet progressed (censored). 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,

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

where f is probability density function and F is the cumulative distribution function, survival function gives us 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 (given in Eqn. (1)) 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 and 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/event times (not censored), d_j is the number of failures/events 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, resulting in a *step* function (the characteristic feature of KM estimate).

2.2. Differential Privacy

Releasing any form of data (raw data, function estimates, derived statistics, etc.) can potentially leak sensitive information (Dinur and Nissim, 2003). Differential privacy (Dwork et al., 2006), a *de facto* standard for providing provable privacy guarantees provides us with the method to quantify such information leakage. Differential privacy is based on the concept of *neighbouring* datasets, that is

Definition 1 (Neighbouring datasets (Dwork et al., 2006)) *Two datasets D, D' are said to be neighbouring if*

$$\exists\ i \in D\ \ s.t.\ D \backslash i = D' \tag{3}$$

which means that D and D' are neighboring datasets if they only differ in any one user.

Definition 2 (Differential privacy (Dwork et al., 2006)) *A randomized mechanism $\mathcal{M} : D^n \to \mathbb{R}^d$ preserves (ϵ, δ)-differentially privacy if for any pair of databases $(D, D' \in D^n)$ such that $d(D, D') = 1$, and for all sets S of possible outputs:*

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in S] + \delta \tag{4}$$

The definition guarantees that it is information-theoretically impossible for an adversary to infer whether the input dataset to the mechanism $\mathcal{M}$ is D or D' (where D, D' are neighboring datasets, that is, $d(D, D') = 1$) beyond a certain probability. The probability is a multiplicative factor of e^{ϵ}. We can see that by making ϵ smaller, we can make the probability small, leading to a strong degree of *plausible deniability* for an individual's presence or absence in the dataset. The definition above allows the *relaxation* of strict privacy guarantees by an additive factor of δ. As is clear from the definition, smaller (ϵ, δ) provide stronger privacy guarantees. When $\delta = 0$, we have pure-ϵ differential privacy.

Differential privacy has many interesting properties, here we briefly introduce the most useful for our use-case. That is the *post-processing*. The post-processing theorem states that differential privacy is immune to post-processing. That is, any function acting solely on the output of a differentially private mechanism is also differentially private, formally

Theorem 3 (Post processing (Dwork and Roth, 2014)) *Let $\mathcal{M} : D^n \to \mathbb{R}$ be a randomized mechanism that is (ϵ, δ)-differentially private. Let $f : \mathbb{R} \to \mathbb{R}'$ be a deterministic function. Then $f \circ M : D^n \to \mathbb{R}'$ is (ϵ, δ)-differentially private.*

This result is directly used in our proposed model. Where after adding noise to our main quantity of interest, we claim that any estimates derived solely from the differentially private quantity are differentially private.

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 unique failure 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/event up to that time point), and we have the number of subjects experiencing the event at that time point (number of progressions/events), which we denote as d_j.

Specifically, we define a function $f(D) \to M$ that takes a dataset D as an input, where each row of D represents an individual with three attributes (t_j, d, r), that are the unique timepoints t_j, an indicator for event d (1/0), and an indicator for being at risk r (1/0). Output of $f(D)$ is a *partial* matrix $M^{k\times 2}$, which stores the results of sum queries on D for d_j for the time $t_j; j \in [1, \cdots, k]$, and for r_1 for the time t_1. That is, M has the data on the number of events (d_1) and the number at risk (r_1) for t_1, and for the rest of the time points $(t_j, j \in [2, \cdots, k])$, we only have the data on the number of events (d_j). We use this initial setup to ensure our privacy guarantees hold, as if we do not use the partial matrix M to start but it's full version with $d_j, r_j; j \in [1, \cdots, k]$, we can encounter scenarios where a

specific person is at risk for multiple time-points (an extreme example is of a person starting the study and never experiencing the event till the maximum followup, leading to their presence in r_j for all time points), leading to *large* sensitivity, and hence an *extremely noisy* result, we avoid this with our partial matrix setup. After creating M, using the derived L_1 sensitivity ($\mathcal{S}$) (details in Section 3.2), 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. Adding Z to M ($M' = M + Z$) guarantees that M' is differentially private (formal proof in Section 3.2).

Please note that the matrix M', although differentially private, is still incomplete as we only have the number at-risk for the first time point (r'_1 for t_1, after noise addition). To complete the matrix, we derive the rest of the *at-risk* population using our noisy events (d'_j). That is, to obtain the number at risk (r'_j) for a subsequent time point (t_j), we subtract the number of events (d'_{j-1}) from the number at risk (r'_{j-1}) for the previous timepoint. This approach is similar to how number at-risk is generally calculated in a non-noisy case, as cases that have experienced an event are no longer at risk for the same event and are removed from the risk set, we use the noisy number of events (d'_j) to ensure our privacy guarantees hold. We present our method succinctly as Algorithm 1 followed by a detailed discussion.

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

procedure DP($\hat{S}(t)$)
  $f(D) \rightarrow M$ ▷ Create a partial matrix M; $[r_1, d_j] \in M$; for every t_j
  $M' = M + Lap(\mathcal{S}/\epsilon)$; $[r'_1, d'_j] \in M'$
  for $j = 2, \cdots, k$ **do**
    $r'_j = r'_{j-1} - d'_{j-1}$
  end for
  $\hat{S}'(t) = \prod_{j:t_j \leq t} \frac{r'_j - d'_j}{r'_j}$
  return $\hat{S}'(t)$
end procedure

3.1.1. Discussion

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 simplistic approach serves a crucial two-fold role. This boosts the accessibility of our differentially private version (it can be implemented using any readily available software package), and aids in ensuring that many other required and reported statistics with the survival function (test statistics, confidence intervals, etc.) are differentially private without spending any extra privacy budget (details follow). Also, in our method, the required changes for differential privacy come with no computational overhead compared to the original estimation (our method is equally computationally cheap).

An important observation is that with current Algorithm 1, using M' for estimating the survival function, we might have scenarios where d'_j or rarely r'_j are negative, leading to the *non-monotonic* behavior of the differentially private survival function. We fix this issue as follows: After completion of M', we check to ensure that any noisy values are not violating

our data integrity constraints (i.e. $r'_j, d'_j < 0$), if they are, we replace such values by 0[3]. This extra step does not require any additional privacy budget and it does not violate our privacy claims, as it is a standard case of *post-processing* in differential privacy, in spirit similar to label smoothing (Wang et al., 2016) or enforcing data integrity constraints (Flaxman, 2019). Another observation is that the algorithm 1 applies to the scenarios in the absence of interval censoring, that is, subjects either have the event or participate to the end of study (up to the maximum follow-up time). However, in many scenarios subjects leave study unexpectedly (moved out of state/country, did not wish to continue/withdrew, etc.) before completion, for such cases our Algorithm 1 can be easily extended, where M is created with $[r_1, d_j, c_j]$, where c_j are the number censored at a time point j. The number at risk completion after noise addition now involves c'_{j-1}, that is, to get r'_j we subtract the sum of noisy events and noisy censored at time $j-1$ from the number at risk at time $j-1$. All privacy guarantees including sensitivity calculations remain the same in case of interval censoring.

Next, we provide the formal privacy guarantees and further details on how our proposed 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 ($\mathcal{S}$) for our method.

Lemma 4 *L_1 sensitivity ($\mathcal{S}$) of $f(D)$ is two.*

Proof As $f(D)$ outputs result of sum queries for d_j and r_1 for t_j, changing one single individual can change the counts (sum) by at most two (that is being in at-risk group and having an event). ■

Theorem 5 *Algorithm 1 is ϵ-differentially private.*

Proof Proof of the differential privacy of M' follows from an instantiation of the Laplace mechanism from Dwork and Roth (2014) with the sensitivity defined in Lemma 4.

As our function estimation ($\hat{S}'(t)$) 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 and Roth, 2014). ■

4. Extending to Other Estimates

As mentioned in the introduction and the previous section, 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 among patient groups, etc. Here we formally define the extensions with their privacy guarantees.

3. Once the at-risk population is 0, we do not consider any future time points.

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 the 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.

4.1.1. 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 6 *Greenwood's (Greenwood et al., 1926) linear point-wise confidence intervals for $\hat{S}'(t)$ are ϵ-differentially private.*

Proof

Greenwood's formula for the confidence intervals is given as

$$\hat{S}(t) \pm z_{1-\alpha/2}\sigma_S(t) \tag{5}$$

where

$$\sigma_s^2(t) = \hat{V}[\hat{S}(t)] \tag{6}$$

and

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)} \tag{7}$$

Replacing by their respective differentially private counterparts from Algorithm 1.

$$\hat{V}'[\hat{S}(t)] = \hat{S}'(t)^2 \sum_{t_j \leq t} \frac{d'_j}{r'_j(r'_j - d'_j)} \tag{8}$$

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 and Roth, 2014). ■

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.

4.1.2. Hypothesis Tests

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 5), the hypothesis testing using the Logrank test is differentially private.

Theorem 7 *Hypothesis test for comparing two groups with the log-rank test for $\hat{S}'(t)$ is ϵ-differentially private.*

Proof For comparing two groups, the log–rank test statistic is formed using the sum of the observed minus expected counts over all failure times for one of the two groups divided by the variance of the summed observed minus expected score (Kleinbaum and Klein, 2010), below we consider group 1.

$$Z = \frac{\sum_{j=1}^{k}(O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{k} V_j}} \tag{9}$$

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

$$E_{1j} = d_j \frac{r_{1j}}{r_j} \tag{10}$$

and V_j is the variance, given as

$$V_j = \frac{r_{1j} r_{2j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)} \tag{11}$$

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

$$V'_j = \frac{r'_{1j} r'_{2j} d'_j (r'_j - d'_j)}{r'^2_j (r'_j - 1)} \tag{12}$$

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 and Roth, 2014). ■

Under the null hypothesis (there is no overall difference between the two survival curves), the log–rank statistic (Z) is approximately chi-square with one degree of freedom. Inference for group comparison can be made by using P-value determined from tables of the chi-square distribution.

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. 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. For complete understanding of competing risks and cumulative incidence, please see Kalbfleisch and Prentice (2011).

Here we show that using Algorithm 1, we can easily extend differential privacy to the competing risk scenarios. However, as our database D now involves an additional term (indicator d_k for event of type k, that is, whether it was the event of interest (of type k) or not), we design D such that when there is "1" for event of interest (for d_k), there is "0" for d, that is, the event is captured either in d (when not of type k) or in d_k (of type k), never in both. We do so to preserve our sensitivity as we will show that we can easily derive complete d using partial d and d_k. For now, our sum queries for time $t_j, j \in [1, \cdots, k]$ include d_j, d_k, and r_1. Same as earlier, we proceed to add noise to the queries with our sensitivity from Lemma 4, which still holds as changing one individual can change the query output by at most two (that is, being in at-risk group r_1, or being in either d or d_k). We then get our complete d required for estimating the cumulative incidence by summing the noisy versions of d and d_k.

Theorem 8 *Competing risk cumulative incidence using our method is ϵ-differentially private.*

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

$$\hat{I}_k(t) = \sum_{j:t_j<t} \hat{S}(t_j)\frac{d_{jk}}{r_j} \tag{13}$$

where d_{jk} is the number of events of type k at time $t_{(j)}$ and $\hat{S}(t_j)$ is the standard Kaplan-Meier estimator to time $t_{(j)}$.

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

$$\hat{I}_k(t)' = \sum_{j:t_j<t} \hat{S}(t_j)'\frac{d'_{jk}}{r'_j} \tag{14}$$

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

Further statistics associated with the cumulative incidence such as the confidence intervals and hypothesis tests, etc. that directly depend on the quantities made differentially private using Algorithm 1 can be similarly argued to be differentially private.

4.3. Nelson-Aalen's Estimate of the Hazard Function

Analogous to the KM estimator of the survival function, another important non-parametric estimator that is often used is the Nelson-Aalen estimator (Nelson, 1969; Aalen, 1978) of the cumulative hazard. Nelson-Aalen estimator estimates the hazard at each time point and has a nice interpretation as the expected number of deaths in $(0, t_j]$ per unit at risk. Although hazard function can be derived using it's relationship with the survival function ($\hat{A}_t = -\log(\hat{S}_t)$), for which we can directly argue differential privacy using our estimate of $\hat{S}'(t)$, there are certain scenarios where we explicitly need to use the Nelson-Aalen estimator or require it for deriving "other" estimates, such as the Flemming-Harrington's estimate of the survival function, etc. Hence, below we prove that using Algorithm 1, we can easily guarantee differential privacy of Nelson-Aalen's estimator.

Theorem 9 *Following Algorithm 1, Nelson-Aalen estimator of the hazard function is differentially private.*

Proof Nelson-Aalen's estimator is given as

$$\hat{A}_t = \sum_{t_j \leq t} \frac{d_j}{r_j} \tag{15}$$

replacing d_j and r_j with their noisy counterparts from Algorithm 1

$$\hat{A}'_t = \sum_{t_j \leq t} \frac{d'_j}{r'_j} \tag{16}$$

where $\hat{A}'_t$ is now differentially private by the post-processing property of differential privacy. ■

Similar to the survival function, "other" statistics associated with the Nelson-Aalen estimator can be argued to be differentially private following Algorithm 1.

5. Empirical Evaluation

Here we present the empirical evaluation of our method on eleven real-life clinical datasets (nine for evaluating KM and two for competing risk) of varying properties. We start with the dataset description for our main comparison.

5.1. Datasets

Nine real-life clinical datasets with time to event information are used to evaluate our proposed method for the KM estimate[4]. Dataset summary is provided in Table 1 followed by further dataset-specific details (dataset properties, pre-processing, group comparison details for hypothesis tests, etc.).

4. We use two additional datasets for competing risk evaluation in Section 5.4

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

Dataset	Observations	Events
Cancer	228	165
Gehan	42	30
Kidney	76	58
Leukemia	23	18
Mgus	1384	963
Myeloid	646	320
Ovarian	26	12
Stanford	184	113
Veteran	137	128

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 survival amongst 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 for the recurrence times to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment (McGilchrist and Aisbett, 1991). Time is converted into months and groups compared are males and females.

4. Leukemia: The dataset pertains to survival in patients with Acute Myelogenous Leukemia (Miller Jr, 2011). Time is converted into months and groups compared are the patients receiving maintenance chemotherapy vs no maintenance chemotherapy.

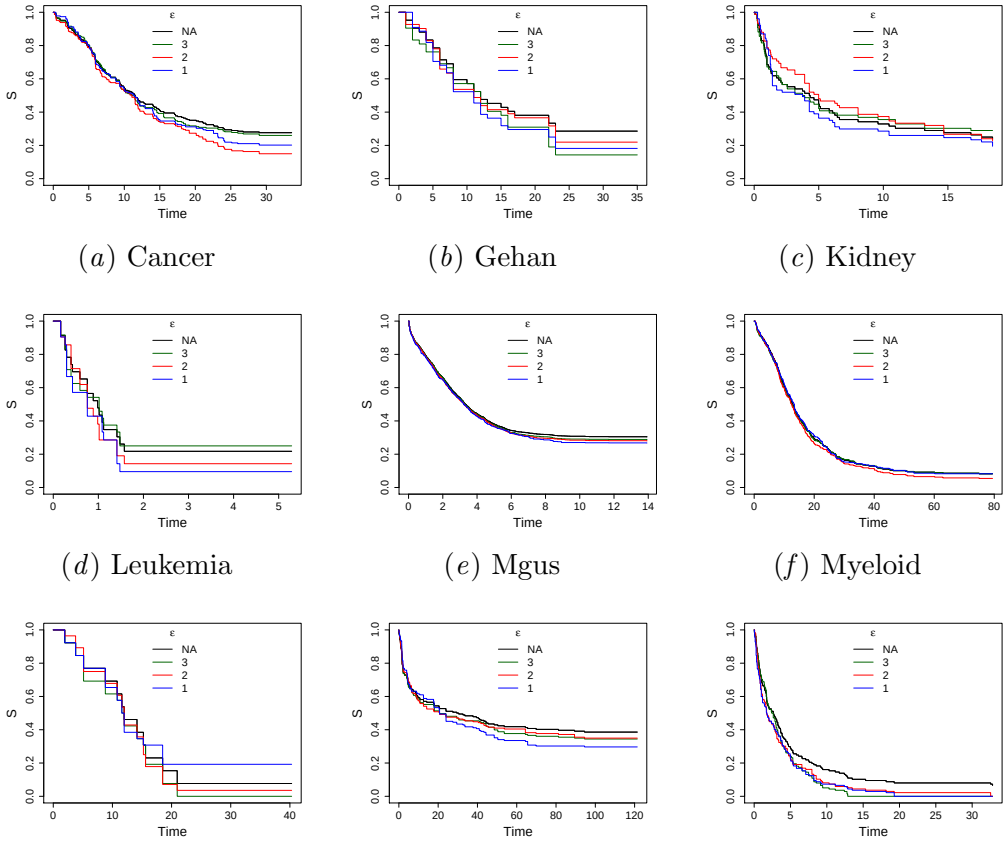
5. Mgus: This dataset is about natural history of subjects with monoclonal gammopathy of undetermined significance (MGUS) (Kyle, 1993). Time is converted into months and groups compared are males and females.

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 is the Stanford Heart Transplant data (Escobar and 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 and Prentice, 2011). Time is converted into months and groups compared are the treatment arms.



(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 $\epsilon = 3$, the orange line is the differentially private estimate with $\epsilon = 2$, and the blue 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.

5.2. Setup and Comparison

To ensure thorough evaluation of our proposed method, we use varying settings for the privacy budget ϵ ($\epsilon \in [3, 2, 1]$). Being a "non-trainable" model, there are no train/test splits and results are reported on the complete dataset. 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[5].

As there is no current method for producing differentially private estimates of the survival function. We compare our approach to the original, gold-standard of the "non-private" estimation. 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 provides credibility to our claim of high utility and will encourage practitioners to adopt our method for practical use.

5.3. Main Results

Now we present the outcome of our evaluation of differentially private KM estimation 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. Estimating Survival Function

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 to no utility loss. As expected, estimation deteriorates with decreased privacy budget ($\epsilon \in [2, 1]$, orange and blue lines respectively). This is intuitive as when the privacy budget decreases, the noise scale required to preserve differential privacy increases, leading to *noisier* estimates.

An observation worth making is that as the dataset size gets smaller (such as ovarian, Leukemia, etc.), the utility of our differentially private estimation gets worse. Which is because from the differential privacy point of view, to protect an individual's privacy in a small dataset, we 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. When tested for statistical differences, we found that all privacy preserving estimates (with $\epsilon \in [3, 2, 1]$) were not statistically-significantly different from the original, non-noisy estimate[6].

5.3.2. Median Survival and Associated Confidence Intervals

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 the high privacy regime. For the performance of our method, we see a similar trend as we saw with results in Figure 1, where our precision increases with increasing dataset size, an acceptable trade-off for individual-level privacy protection.

5. Link removed to respect double blind review process.
6. Using the logrank test with statistical significance set at 0.05 level

Table 2: Median Survival with associated confidence intervals. 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)				
Dataset	$\epsilon = 3$	$\epsilon = 2$	$\epsilon = 1$	No Privacy
Cancer	11.5 (9.5, 14.1)	11.4 (9.4, 12.2)	11.1 (9.4, 11.9)	11.5 (9.5, 14.1)
Gehan	12.0 (7.0, 16.0)	11.0 (7.0, 18.0)	11.0 (7.0, 22.0)	12.0 (8.0, 14.1)
Kidney	3.9 (1.9, 6.6)	4.9 (3.9, 8.0)	4.8 (3.5, 8.4)	4.3 (1.4, 6.2)
Leukemia	1.0 (0.3, 1.5)	0.8 (0.2, 1.0)	0.7 (0.1, 1.1)	0.9 (0.4, 1.4)
Mgus	3.3 (3.1, 3.5)	3.3 (3.0, 3.5)	3.2 (3.0, 3.5)	3.3 (3.1, 3.6)
Myeloid	12.6 (11.9, 13.5)	12.5 (11.7, 13.4)	13.4 (12.2, 13.8)	12.7 (11.9, 13.8)
Ovarian	12.0 (8.9, 15.6)	11.8 (8.8, 15.2)	11.6 (5.1, 15.2)	11.9 (8.8, 15.2)
Stanford	26.3 (15.8, 49.3)	24.7 (12.8, 47.3)	21.4 (18.7, 33.6)	30.5(12.5, 48.5)
Veteran	2.6 (1.7, 3.0)	1.7 (1.2, 2.8)	1.7 (1.1, 2.7)	2.6 (1.7, 3.4)

5.3.3. Test Statistic

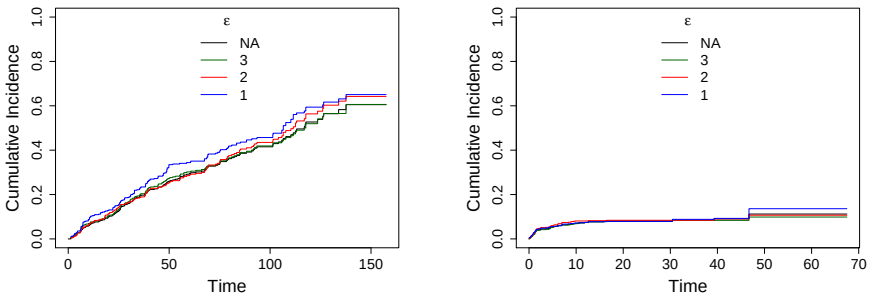
For the test statistic (obtained from comparing the survival distribution of different groups in the dataset, group details provided in the Section 5.1), in Table 3, we observe that our differentially private estimation performs at par with the original "non-noisy" estimation, even for the high privacy regime ($\epsilon \in [2, 1]$). 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.

Table 3: 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 good utility with strong privacy guarantees.

Test Statistic (Z)				
Dataset	$\epsilon = 3$	$\epsilon = 2$	$\epsilon = 1$	No Privacy
Cancer	11.1	11.8	11.9	10.3
Gehan	17.5	17.6	28.1	16.3
Kidney	7.4	8.4	21.9	6.9
Leukemia	2.9	2.4	2.5	3.4
Mgus	7.2	6.9	5.9	9.7
Myeloid	9.2	9.1	10.7	9.6
Ovarian	0.8	1.2	2.4	1.1
Stanford	6.2	6.9	8.0	6.6
Veteran	0.2	0.3	1.1	0.02

5.4. 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 and 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).



(*a*) PBC (*b*) Transplant

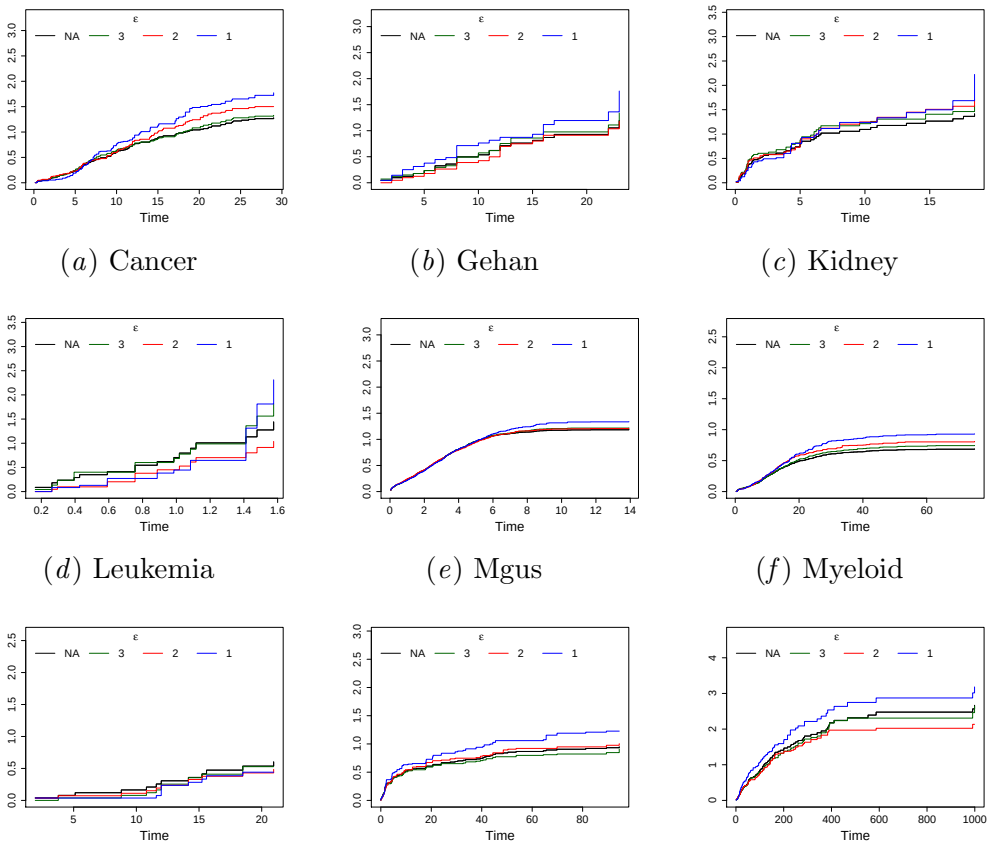
Figure 2: Extending differentially private estimation to competing risk cumulative incidence (cumulative incidence is the opposite of survival function, so the plots go upward). Black is the original, unperturbed estimate. Green is with $\epsilon = 3$, orange is with $\epsilon = 2$, and blue is with $\epsilon = 1$. We can see that our method does a good job of estimating competing risk cumulative incidence while providing strong privacy guarantees.

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.

5.5. Nelson-Aalen Estimate

For evaluating the performance of our proposed differentially private Nelson-Aalen's estimator of the hazard function, we use the main nine datasets. Please note that similar to the competing risk cumulative incidence, being a "risk" estimate, the value of the cumulative hazard estimate increases over time, hence it has an "upward" curve compared to the "downward" curve for the survival estimate.

Figure 3 shows the results for all nine datasets. Our differentially private estimate performs extremely well, similar to our main comparison, where we can see that our estimation provides good utility, even at high privacy regimes. Also, similar to our main comparison, all differentially private estimates are not statistically-significantly different from the original, non-noisy estimates.



(a) Cancer (b) Gehan (c) Kidney

(d) Leukemia (e) Mgus (f) Myeloid

(g) Ovarian (h) Stanford (i) Veteran

Figure 3: Differentially private estimation of the Nelson-Aalen estimator, followup time is on the X-axis and the hazard estimate is on the Y-axis. The black line is the original function estimate, the green line is the differentially private estimate with $\epsilon = 3$, the orange line is the differentially private estimate with $\epsilon = 2$, and the blue line is the differentially private estimate with $\epsilon = 1$. We observe that our differentially private version provides good utility while protecting an individual's privacy.

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 principles, Nguyên and 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;

Esteban et al., 2017; Triastcyn and 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 and Limitations

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 and reported statistics such as the associated confidence intervals, test statistics, and to other estimates such as the competing risk cumulative incidence and the Nelson-Aalen estimate of the hazard function. With extensive empirical evaluation on eleven real-life datasets, we have shown that our proposed method provides a good privacy-utility trade-off. And to aid in rapid adaptation, we have made the source code publicly available.

However, as with any new method, our method has some limitations. As observed during empirical evaluation and as discussed in Section 3.1.1, for smaller datasets there can be scenarios where we have to truncate the function estimation when our *noisy* at-risk population reaches zero. The truncation may lead to biased estimates where the noisy estimate is missing information from events at the tail-end. This is in addition to overall noisier estimation due to small sample size. This phenomenon restricts the end-user to rely on *weaker* privacy guarantees for small datasets (using larger ϵ). The limitations shape our future work, where we would like to investigate the impact of considering different privacy definitions, mechanisms, and noise distributions to minimize the *utility gap* when transitioning from non-private to differentially private regime.

Acknowledgments

This research is in part supported by a CGS-D award and a discovery grant from Natural Sciences and Engineering Research Council of Canada.

References

Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 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*, page 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.

Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210. ACM, 2003.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):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, pages 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*, pages 507–528, 1992.

Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

Abraham D Flaxman. Empirical quantification of privacy loss with examples relevant to the 2020 us census. 2019.

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)*, pages 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*, pages 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.

David G Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 2010.

Robert A Kyle. "benign" monoclonal gammopathy—after 20 to 35 years of follow-up. In *Mayo Clinic Proceedings*, volume 68, pages 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*, pages 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*, pages 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.

Yue Wang, Xintao Wu, and Donghui Hu. Using randomized response for differential privacy preserving data collection. In *EDBT/ICDT Workshops*, volume 1558, 2016.

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*, pages 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.