Privacy-Preserving Models for Comparing Survival Curves Using the Logrank Test[*]

Tingting Chen   Sheng Zhong
Computer Science and Engineering Department
State University of New york at Buffalo
Amherst, NY 14260, U. S. A
Email : {tchen9, szhong}@buffalo.edu

Abstract

The incorporation of electronic health care in medical institutions will benefit and thus further boost the collaborations in medical research among clinics and research institutions. However, privacy regulations and security concerns make such collaborations very restricted. In this paper, we propose privacy preserving models for survival curves comparison based on logrank test, in order to perform better survival analysis through the collaboration of multiple medical institutions and protect the data privacy. We distinguish two collaboration scenarios and for each scenario we present a privacy preserving model for logrank test. We conduct experiments on the real medical data to evaluate the effectiveness of our proposed models.

Keywords: Survival Curves; Privacy Preservation; Logrank Test

1 Introduction

With the development of information technology, there is an increasing need to incorporate electronic health record (EHR) in medical institutions [1]. The availability of EHRs is believed to be able to improve the health care efficiency and quality that the patients receive. Moreover, because of using EHR instead of paper-based records, hospitals can store and manage more health care data than ever before. Consequently, it will benefit the development of more advanced clinical computer-based tools that help diagnosis and research. Especially, if multiple medical institutions can integrate their electronically stored health care data, with this substantial amount of data, better models with higher accuracy can be built to assist clinical treatment and medical research.

[*] Correspondence Author: Sheng Zhong, Department of Computer Science and Engineering, State University of New York at Buffalo, Amherst, NY 14260, U. S. A. Phone: +1-716-645-4752 Fax: +1-716-645-3464. Email: szhong@buffalo.edu

Survival analysis [2] is an important statistic tool often used in clinic trial to provide assessment of benefit and risk. With the collaboration of multiple medical institutions, researchers or doctors can build better survival analysis models, especially survival function comparison models. Here we illustrate two different scenarios as examples. The first scenario is that in a hospital, a new radiotherapy treatment is performed to a group of pancreatic cancer patients. The doctors in the hospital can observe the survival events (death or cancer recurrence) of these patients and draw a survival curve for this new treatment. They want to compare this survival curve with other treatments to justify its effectiveness and advantage. Luckily, a medical research institution holds survival data of other treatment trials for pancreatic cancer with trial participants of similar background. Clearly the collaborative data exchange between the hospital and the research institutions will be beneficial for the result comparison. In the second scenario, three institutions are all studying the performance of a new medicine for stroke on patients of different ages. They want to build and compare the survival curves for different age intervals. However, the trial participants in any one of the three institutions are not sufficient to obtain results with high accuracy. If they can conduct survival curves comparison based on the trial participants from all of them, it will significantly increase the result accuracy.

However, sharing medical data is well-known to be restricted because of privacy and security concerns. According to the privacy rules of Health Insurance Portability and Accountability Act (HIPAA) [3], the privacy of patients must be protected and it is illegal for research institutions and hospitals to distribute patient's medical data without appropriate privacy preservation. On the other hand, Medical researchers are reluctant to share their data with others even if it is already anonymized, due to the concern of possibility that their data could be misused or misinterpreted. For instance, in Dartmouth College neuroscientist found it difficult to encourage the sharing of brain imaging data [4]. In the two scenarios above, the privacy concern also exists which impedes the process of collaboration between medical institutions. Therefore, we need to develop new models for survival curves comparison that can protect the privacy of patients and relieve the data security concern of the researchers or doctors.

In this paper, we propose novel privacy preserving models for logrank test, which is a standard comparison test of survival curves. In particular, for each of the two collaboration scenarios we mentioned above, we design one privacy-preserving logrank test model. In the rest of this paper, we call the first scenario group partition, meaning each institution holds a survival curve for a entire group of participants. We call the second scenario sample partition, meaning each institution holds the survival data of some (but not all) participants in each group. Our goal is that for each of the collaboration scenario, our proposed logrank test model can learn the comparison result of survival curves built on the data from all medical institutions, even without looking at the original survival data from other medical institutions. We utilize a cryptographic tool, secure sum[5], in our models. In this way, the privacy of medical data is protected. As far as we know, it is the first work on building privacy preserving models for survival curves comparison using logrank test. We preform experiments on real medical data to show the effectiveness of our proposed models.

2 Methods

In this section, we first review the logrank test for comparing survival curves. Then we describe our two privacy preserving models for the logrank test. The first model enables the privacy preserving comparison of survival curves in the group partition scenario. Then we present the second privacy preserving model which preserves the privacy in comparing the survival curves in the sample partition scenario.

2.1 Overview of Logrank Test

Suppose we have n groups of individuals. Logrank test [2] is a statistical hypothesis test, where the hypothesis is that the n groups have the same survival distribution, i.e., for each group the probability of occurring the event (e.g., death) at each time point is the same. In particular, we divide the time into m intervals. Let n_{kj} be the number of individuals that are alive in group k at the beginning of time interval j. Let d_{kj} be the number of events occurring in group k in interval j. n_j and d_j are defined as Eq. (1) and Eq. (2) respectively.

$$n_j = \sum_{k=1}^{n} n_{kj} \tag{1}$$

$$d_j = \sum_{k=1}^{n} d_{kj} \tag{2}$$

The test statistic is calculated as

$$Z = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}, \tag{3}$$

where O_k represents the number of observed deaths in group k, i.e.,

$$O_k = \sum_{j=1}^{m} d_{kj}. \tag{4}$$

E_k is the expected number of deaths in group k, i.e.,

$$E_k = \sum_{j=1}^{m} \frac{n_{kj} d_j}{n_j}. \tag{5}$$

A smaller test statistic Z suggests a higher probability that the hypothesis is true.

2.2 Privacy for Each Party

As mentioned above, it is often the case that the survival data for several groups are distributed in different places, e.g., medical research institutions and clinics. These organizations or parties want to compare their survival curves using logrank test but each of them is not willing to reveal its own survival data to other parties. We distinguish the privacy of each party for the two collaboration scenarios, i.e., the group partition and the sample partition.

- **The Group Partition**
  In the group partition scenario, each party holds the survival data collected from the group of patients that this party has: number of events occurring in each time slot and the number of surviving individuals at the beginning of each time interval. Without loss of generality, we assume that party k holds the survival data of group k. In our proposed privacy-preserving logrank test model, we aim to for each party k protect the information n_{kj} and d_{kj} $(\forall j)$ from other parties than k and meanwhile correctly compute the logrank test statistic. The group partition scenario is illustrated in Fig. 1.

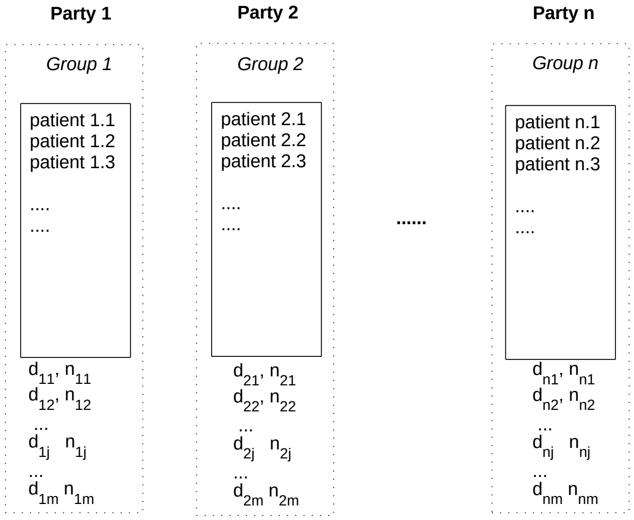

Figure 1: The Group Partition Scenario.

- **The Sample Partition**
  In the sample partition scenario, each party holds the survival data for some participants in each group. Formally, $\forall\ k, j$ each party i holds its survival data n^i_{kj} and d^i_{kj}, which are collected for time interval j from the patients in group k that party i has. Each party i wants to keep n^i_{kj},

and d^i_{kj} private and build logrank model using n_{kj} and d_{kj}, such that $n_{kj} = \sum_i n^i_{kj}$ and $d_{kj} = \sum_i d^i_{kj}$. Figure 2 shows the sample partition scenario.

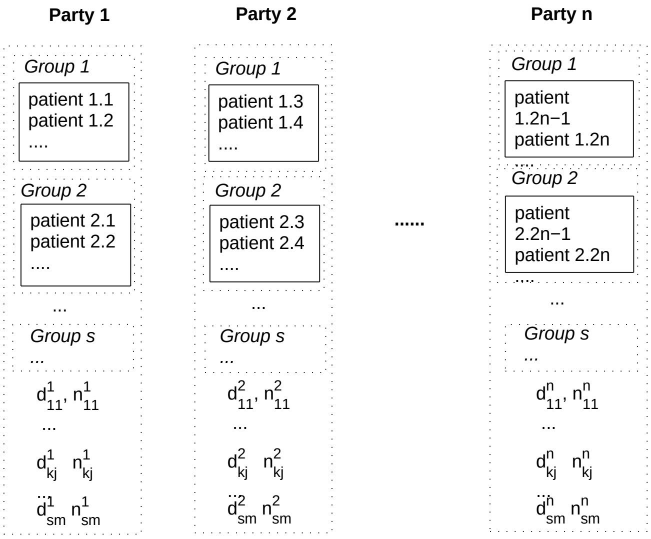

Figure 2: The Sample Partition Scenario.

2.3 Privacy-Preserving Logrank Test Model for the Group Partition Scenario

Assume that there are s parties ($s \geq 3$), each of which holds the survival data for one group of patients. The s parties want to jointly perform the logrank test, (i.e., computing the value Z), without revealing their private data to each other.

From Eq. (3) we know that if party k wants to compute Z, it needs to be able to compute E_k first, but from Eq. (3) (1) and (5), we know that

$$E_k = \sum_{j=1}^{m} \frac{n_{kj} \sum_k d_{kj}}{\sum_k n_{kj}}, \tag{6}$$

which requires $d_{k'j}$ and $n_{k'j}$ where $k \neq k'$. Moreover, in order to compute Z, party k needs to know $\sum_{k' \neq k} \frac{(O'_k - E'_k)^2}{E'_k}$ to obtain the sum.

In our model, we utilize a randomization based secure computation tool, secure sum [5], to tackle these two challenges. Now we first describe how to securely compute E_k for each k, without knowing $d_{k'j}$ and $n_{k'j}$ where $k \neq k'$. Then we present our complete privacy-preserving logrank test model.

Suppose that all participating parties are numbered as parties $1 \cdots s$. Our method of computing E_k for each party k is summarized in Fig. 3. The idea of our method is that to compute each d_j (resp. n_j) $\forall j$, party 1 generates a random number with the same range as that of d_{ij} and n_{ij} and adds its local value d_{1j} (resp. n_{1j}) to the random number before passing it to the next party. In this way the actual values of d_{1j} and n_{2j} are hidden behind the random numbers. Similarly, every other party adds its local value to the sums that it receives and sends the new sums to the next party. Finally, party s sends $R_{j1} + \sum_{k=1}^{s-1} d_{kj}$, and $R_{j2} + \sum_{k=1}^{s-1} n_{kj}$ back to party 1, and party 1 subtracts the two random numbers respectively and obtains d_j and n_j. After party 1 sends $\frac{d_j}{n_j}$ to all other parties, each party k can compute the value of E_k.

```
For interval $1 \leq j \leq m$,
    (Step 1): Party 1 generates two random numbers $R_{j1}$ and $R_{j2}$ (s.t., $R_{j1} \in [0, D)$, $R_{j2} \in [0, D)$)[a] and sends $R_{j1} + d_{1j}$ and $R_{j2} + n_{1j}$ to Party 2.
    (Step 2): For each party p, s.t. $2 \leq p \leq s - 1$,
        Party $p$ receives $R_{j1} + \sum_{k=1}^{p-1} d_{kj}$, and $R_{j2} + \sum_{k=1}^{p-1} n_{kj}$. Party $p$ computes $R_{j1} + \sum_{k=1}^{p-1} d_{kj} + d_{pj}$ and $R_{j2} + \sum_{k=1}^{p-1} n_{kj} + n_{pj}$ and sends them to party $p + 1$.
    (Step 3): Party $s$ receives $R_{j1} + \sum_{k=1}^{s-1} d_{sj}$, and $R_{j2} + \sum_{k=1}^{s-1} n_{kj}$. Party $s$ computes $R_{j1} + \sum_{k=1}^{n-2} d_{kj} + d_{sj}$ and $R_{j2} + \sum_{k=1}^{s-1} n_{kj} + n_{sj}$ and sends them to party 1.
    (Step 4): Party 1 receives $R_{j1} + \sum_{k=1}^{s-1} d_{kj}$, and $R_{j2} + \sum_{k=1}^{s-1} n_{kj}$. Party 1 substracts $R_{j1}$ and $R_{j2}$ from the two received numbers respectively, and computes $\frac{d_j}{n_j} = \frac{\sum_k d_{kj}}{\sum_k n_{kj}}$. Party 1 sends $\frac{d_j}{n_j}$ to all other parties.
End For

Each party $k$ (s.t., $1 \leq k \leq s$) computes $E_k = \sum_{j=1}^{m} n_{kj} \frac{d_j}{n_j}$.
```

[a] Here D is the range of d_{ij} and n_{ij}, $\forall i, j$.

Figure 3: Privately Computing E_k for each party k.

After each party k gets E_k, it can easily compute $\frac{(O_k - E_k)^2}{E_k}$ because obtaining O_k does not require the survival data from other parties. As the final step of our privacy-preserving logrank test model, we need to compute $\sum_k \frac{(O_k - E_k)^2}{E_k}$. Since $\frac{(O_k - E_k)^2}{E_k}$ reveals the information of how much the survival curve of group k is different from other groups, we also need to keep it private when calculating Z. Again we use the idea of secure sum. Party 1 generates a random number R and adds $\frac{(O_1 - E_1)^2}{E_1}$ to it before passing it to the next party. When each party k has added the $\frac{(O_k - E_k)^2}{E_k}$, party 1 again receives $R + \sum_{k=1}^{s} \frac{(O_k - E_k)^2}{E_k}$. Party 1 subtracts R and sends the value of Z to other parties.

We notice that in our privacy-preserving Logrank test model, party 1 meed to generate random numbers. In the implementation, we use the pseudo random number generator function in the GNU C Library [6]. Applying pseudo random number generation algorithm is a standard way to generate random numbers in cryptography.

2.4 Privacy-Preserving Logrank Test Model for the Sample Partition Scenario

Assume that there are $n(n > 2)$ parties want to compare s survival curves in the sample partition scenario. As we have mentioned, in the sample partition scenario, each party i holds its survival data d^i_{kj} and n^i_{kj}. They would like to collaboratively compute Z that can be written as follows

$$Z = \sum_{k=1}^{s} \frac{(\sum_{j=1}^{m} \sum_{i=1}^{n} d^i_{kj} - \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} n^i_{kj} \sum_{i=1}^{n} d^i_j}{\sum_{i=1}^{n} n^1_j})^2}{\sum_{j=1}^{m} \frac{\sum_{i=1}^{n} n^i_{kj} \sum_{i=1}^{n} d^i_j}{\sum_{i=1}^{n} n^1_j}}. \tag{7}$$

As we can see, to compute Z in a privacy preserving way for the sample partition scenario is more complicated than for the group partition scenario. Again we utilize the idea of secure sum. For ease of presentation, we only describe the main steps in this model and skip the details of secure sum computation which is similar to what we have presented above.

First, using the secure sum algorithm, the n parties can securely obtain

$$O_k = \sum_{i=1}^{n} \sum_{j=1}^{m} d^i_{kj} \tag{8}$$

and $\sum_{i=1}^{n} n^i_{kj}$. Similarly, for each j, the n parties can securely compute $\frac{\sum_{i=1}^{n} d^i_j}{\sum_{i=1}^{n} n^i_j}$, where $d^i_j = \sum_k d^i_{kj}$ and $n^i_j = \sum_k n^i_{kj}$ can be computed locally at each party i. With $\sum_{i=1}^{n} n^i_{kj}$ and $\frac{\sum_{i=1}^{n} d^i_j}{\sum_{i=1}^{n} n^i_j}$, now E_k can be computed as in Eq. (9)

$$E_k = \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} n^i_{kj} \sum_{i=1}^{n} d^i_j}{\sum_{i=1}^{n} n^1_j} \tag{9}$$

After obtaining O_k and E_k for each party k, apply the secure sum again to securely compute the sum of $\frac{(O_k - E_k)^2}{E_k}$ for the s groups.

3 Results

In this section, we perform experiments on our privacy preserving logrank models for the two scenarios using real survival data of kidney patients [7].

3.1 Data Description

We use the data in [7] on the recurrence times to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. Catheters may be removed for reasons other than infection, in which case the observation is censored.

For the experiments on our model for the group partition scenario, we divide the survival data into three groups according to the type of disease (Glomerulo Nephritis, Acute Nephritis and Polycystic Kidney Disease) that the kidney patient has, and assume that there are three parties each of whom is holding the survival data for one group of a disease type. Table 1 summarizes the survival data that each party has.

Table 1: Survival data for the group partition scenario

Time to infection	d_{1j}	n_{1j}
0-50	11	18
51-100	0	7
101-150	2	7
151-200	4	5
201-250	0	1
251-300	0	1
301-350	0	1
351-400	0	1
401-450	0	1
451-500	0	1
>500	1	1

(a) Party 1

Time to infection	d_{2j}	n_{2j}
0-50	14	24
51-100	5	10
101-150	2	5
151-200	1	3
201-250	1	2
251-300	0	1
301-350	1	1
351-400	0	0
401-450	0	0
451-500	0	0
>500	0	0

(b) Party 2

Time to infection	d_{3j}	n_{3j}
0-50	3	8
51-100	2	5
101-150	0	3
151-200	2	3
201-250	0	1
251-300	0	1
301-350	0	1
351-400	0	1
401-450	0	1
451-500	0	1
>500	1	1

(c) Party 3

In the experiments on the group partition scenario, we divide the survival data into two groups based on the ages of patients. The two age intervals for forming groups are (20-50] and (50-70]. We assume that there are three parties participating in the survival data comparison. Each of the three parties holds some samples for both the two groups of patients. Table 2 shows the survival data distribution in this scenario.

Table 2: Survival data for the sample partition scenario

	Group 1		Group 2	
Time	d_{1j}^1	n_{1j}^1	d_{2j}^1	n_{2j}^1
0-50	4	7	8	14
51-100	0	3	1	6
101-150	1	3	4	5
151-200	2	2	0	1
201-250	0	0	0	1
251-300	0	0	0	1
301-350	0	0	0	1
351-400	0	0	0	1
401-450	0	0	0	1
451-500	0	0	0	1
> 500	0	0	1	1

(a) Party 1

	Group 1		Group 2	
Time	d_{1j}^2	n_{1j}^2	d_{2j}^2	n_{2j}^2
0-50	6	12	6	9
51-100	2	6	2	3
101-150	1	4	1	1
151-200	1	3	0	0
201-250	0	2	0	0
251-300	0	2	0	0
301-350	1	2	0	0
351-400	0	1	0	0
401-450	0	1	0	0
451-500	0	1	0	0
> 500	1	1	0	0

(b) Party 2

	Group 1		Group 2	
Time	d_{1j}^3	n_{1j}^3	d_{2j}^3	n_{2j}^3
0-50	12	23	4	10
51-100	1	11	2	6
101-150	4	10	0	4
151-200	0	6	3	4
201-250	1	6	1	1
251-300	1	5	0	0
301-350	1	4	0	0
351-400	0	3	0	0
401-450	2	3	0	0
451-500	0	1	0	0
> 500	1	1	0	0

(c) Party 3

Table 3: Intermediate results of privacy preserving logrank model for group partition scenario

Time to infection	d_j	n_j	
0-50	28	50	
51-100	7	22	
101-150	4	15	
151-200	7	11	
201-250	1	4	
251-300	0	3	
301-350	1	3	
351-400	0	2	
401-450	0	2	$E_1 = 18.94$
451-500	0	2	$E_2 = 20.70$
> 500	2	2	$E_3 = 10.36$

3.2 Experimental Results

3.2.1 Results for the Group Partition Scenario

Using the data in group partition scenario shown in Table 1, we apply our privacy preserving logrank model to compare the three survival curves for the three different diseases. The results in the intermediate steps are recorded in Table 3.

The final privacy preserving logrank test statistic is 1.11, which is same with the logrank test result conducted on one site with the data from all the three parties.

3.2.2 Results for the Sample Partition Scenario

For the sample partition scenario, we first perform three logrank tests, each using the data held by one of three parties. Then we conduct our privacy-preserving logrank test model with the collaboration of the three parties. We compare the results to see the difference.

Fig. 4 shows the result of our privacy preserving model for comparing the two groups of patients using the data from all the three parties, and also the result of the basic logrank model using the data from only one party. We can see that if we only use the data from one party, especially party 1, the comparison result is significantly different from the case when we can include more data in the analysis. Therefore our privacy preserving logrank model provides a way to conduct better survival analysis by enabling to securely use more data from different parties.

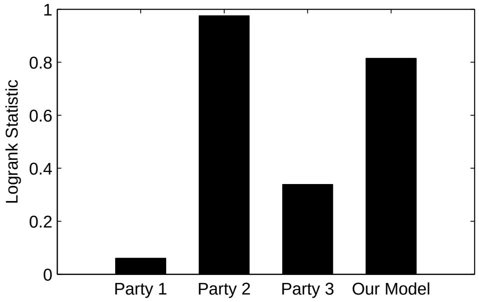

Figure 4: Comparison of logrank test results between our privacy preserving model working on the data from all three parties and the basic logrank model running on the data of one party.

4 Discussion

Our proposed privacy-preserving logrank test models have the following characteristics that are different from previous work.

- They preserve the data privacy for each party without revealing it to others, in the process of survival data analysis using logrank test. Relieving the privacy concerns for the medical institutions is of great importance because it will encourage more cooperations among these institutions on either research or clinical trials. Consequently, better survival data comparisons supported by larger database will become available.
- Our privacy preserving models are accurate, meaning the logrank test results obtained by our models are the same as those obtained by having all data on one site.
- The two models either for the group partition scenario or for the sample partition scenario only require the parties who hold the survival data to participate. In other words, as long as those parties with data use our models, all computation is conducted within those parties and thus no other agencies are needed. Therefore, our models are very practical in its implementation.
- We are utilizing a randomization-based method in our models. As a result our models are much more efficient compared with cryptography-based approaches.

An existing work in data mining community proposed the privacy preserving cox regression in survival analysis [8]. The proposed model was based on linearly projecting the data to a lower dimensional space through an optimal mapping. Different from their model, we focus on another very important function of survival data analysis in medical trials, the survival curves comparison. Furthermore, we did not lose any accuracy in our privacy preserving models.

5 Hardware and software specifications

The models are implemented using GNU C Library. The programs are running in Redhat Linux 7.2 on 2.0GHz computers.

6 Conclusion

In this paper, we propose privacy preserving models for survival curves comparison based on logrank test, in order to perform better survival analysis through the collaboration of multiple medical institutions and protect the data privacy. We distinguish two collaboration scenarios, the group partition scenario and the sample partition scenario. For each scenario we present a privacy preserving model for logrank test. Our experiments on the real medical data to evaluate the effectiveness of our proposed models.

References

[1] Health Information Technology for Economic and Clinical Health Act. Available at: http://waysandmeans.house.gov/media/pdf/111/hitech.pdf

[2] Altman, DG. Practical Statistics for Medical Research. Chapman & Hall. London, 1991. ISBN 0-412-27630-5.

[3] HIPPA, National Standards to Protect the Privacy of Personal Health Information, [Online]. Available at: http://www.hhs.gov/ocr/hipaa/finalreg.html.

[4] Editorial. Whose scans are they, anyway?, Nature, 406 (443), 3 August 2000.

[5] M. Kantarcioglu and C. Clifton, Privacy-preserving distributed mining of association rules on horizontally partitioned data, in The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02), Madison, Wisconsin, June 2 2002, pp. 24-31.

[6] GNU C library, available at: http://www.gnu.org/software/libc/.

[7] McGilchrist and Aisbett, Biometrics 47, 461-66, 1991.

[8] S. Yu, G. Fung, R. Rosales, S. Krishnan, R. B. Rao, Privacy-Preserving Cox Regression for Survival Analysis, in Proceedings of KDD'08, August 2008, Las Vegas, Nevada, USA.