journal homepage: www.intl.elsevierhealth.com/journals/cmpb

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, USA

ARTICLE INFO

Article history:
Received 16 June 2010
Received in revised form
19 April 2011
Accepted 24 April 2011

Keywords:
Survival curves
Privacy preservation
Logrank test

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.

© 2011 Elsevier Ireland Ltd. All rights reserved.

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.

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.

* *Corresponding author*. Tel.: +1 716 645 4752; fax: +1 716 645 3464.
E-mail addresses: tchen9@buffalo.edu (T. Chen), szhong@buffalo.edu (S. Zhong).

0169-2607/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2011.04.004

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.
- **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

Download English Version:

https://daneshyari.com/en/article/468729

Download Persian Version:

https://daneshyari.com/article/468729

Daneshyari.com