Privacy-Preserving Data Sharing and Analysis of Survival Data

Zhe Yuan

Advisors: James Li, PhD, Ming Tan, PhD

Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington, DC

1. Introduction

With the rapid advances of information technology application on healthcare, e.g. cloud computing, big data, etc., more and more institutions are using electronic medical record (EMR) systems and going to paperless, and further store and share the EMR data on the cloud and Internet to maximize its usage. However, due to the nature of the EMR data, privacy is an inevitable issue on data sharing.

1. Electronic medical record is beneficial to patients, physicians, and researchers.
2. Survival analysis is one of the major areas of medical research which needs patient observations.
3. Sharing data directly may violate the privacy of patients according to HIPAA.
4. New analysis approaches are needed for the public good.

Thus, a system that can help those institutions to share their patient related data is urgently needed. This study proposed a solution to protect the privacy of patient in a multi-party data sharing setting.

2. Methods

In a multi-center setting, the process of data sharing follows two cycles: accumulating pseudo data from each center, and deducting noise from the accumulated data (flow chart).

Consider the expected value of pseudo survival time as the same as real survival for any observations;

4 New analysis approaches are needed for the public good.

Since the noise is added to the data by a random number with zero mean and finite variance, there may exist useful information in the pseudo data. Two ways can be used to measure the information in the accumulated pseudo data (step 4 in flowchart) before the second noise deduction procedure in the previous data sharing protocol:

3. Results

From Kaplan-Meier curves in Figure-1, the enciphered data was fully recovered by using sharing protocol. The difference between the pseudo and real data was small, but was getting larger when the variance of noise increases, which is also reflected in the log-rank test in Table-

In this study, we perform experiments on privacy data sharing and utility measurement in pseudo data using real survival data of patients having circulating tumor cells (CTC). The overall survival or last follow up time of each patient are recorded and the CTC measurements at the baseline are also recorded. In the privacy preserving data sharing scenario, the dataset was randomly divided into three groups and the sequence of data sharing was group 1 to group 3, to group 2. For the pseudo data generating procedure, real survival time was replaced by a random variable in Gamma distribution, with mean equal to the real time and the variance equals to 1/100, 1/10, 1, 5, 10, 50 times the real time.

In the utility measuring scenario, the accumulated pseudo data in each of the six experiments were fitted by the cox proportional hazard model, and the hazard ratios for CTC measurement were estimated and compared with the estimation based on the real data.

4. Discussion and Conclusion

A privacy preserving data collaborative analysis protocol was proposed to prevent participated parties from knowing the source of the data in a multi-party setting. Original data can be masked and shared by adding noise, and can be fully recovered using the distinct tracing ID of pseudo data known by each collaborator. Compared with the ordinary survival data analysis models, the proposed method adds Gamma distributed noise tends to modify the survival distribution, but the estimations of hazard ratio did not change by much when the variance of noise is less than the survival time.

Further work may need to be done to extend this sharing protocol. The application of chi-square, lognormal, or other non-negative distributions other than Gamma distribution could be tested in the noise generating process. Since the accuracy of hazard ratio estimation changes with the variance of noise, more studies are needed to find the limit of variance in order to preserve the accuracy in the estimation.

Reference

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

Table-1: Log-rank test to measure the difference between the real survival data and the pseudo data.

<table>
<thead>
<tr>
<th>Hazard ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>real data</td>
<td>2.181</td>
</tr>
<tr>
<td>1/100</td>
<td>2.18</td>
</tr>
<tr>
<td>1/10</td>
<td>2.143</td>
</tr>
<tr>
<td>1</td>
<td>2.143</td>
</tr>
<tr>
<td>5</td>
<td>2.12</td>
</tr>
<tr>
<td>10</td>
<td>2.039</td>
</tr>
<tr>
<td>20</td>
<td>1.965</td>
</tr>
</tbody>
</table>

Table-2: Hazard ratios of survival time stratified by CTC at baseline based on real data and pseudo data in six data sharing experiments with mean equals to the real time and variances at 1/100, 1/10, 1, 5, 10, 50 times the real time.

The estimation of of hazard ratio did not change by much when the variance of noise equals to the survival time, but there was a trend that the estimation of hazard ratio was decreasing in this experiment even though this is not significantly.