

Efficient Differentially Private Methods for a Transmission Disequilibrium Test in Genome Wide Association Studies



Akito Yamamoto, Tetsuo Shibuya

Division of Medical Data Informatics, Human Genome Center,
The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan

Background

· To achieve the provision of personalized medicine, it is very important to investigate the relationship between diseases and human genomes.

However...

- There is a risk of identifying individuals if the statistics are released as they are [4].
- Existing privacy-preserving methods [3] for a transmission disequilibrium test are computational intensive.

Transmission Disequilibrium Test

Number of parents for TDT in one SNP.

		Non-Transmitted Allele		Total
		M_1	M_2	
Transmitted Allele	M_1	a	b	$a + b$
	M_2	c	d	$c + d$
Total		$a + c$	$b + d$	$2n$

$$\chi_{td}^2 := \chi_{td}^2(b, c) = \frac{(b - c)^2}{b + c}$$

Number of families for each (b, c) .

(b, c) in a family	(1, 0)	(0, 1)	(1, 1)	(2, 0)	(0, 2)	(0, 0)
Number of families	n_1	n_2	n_3	n_4	n_5	n_6

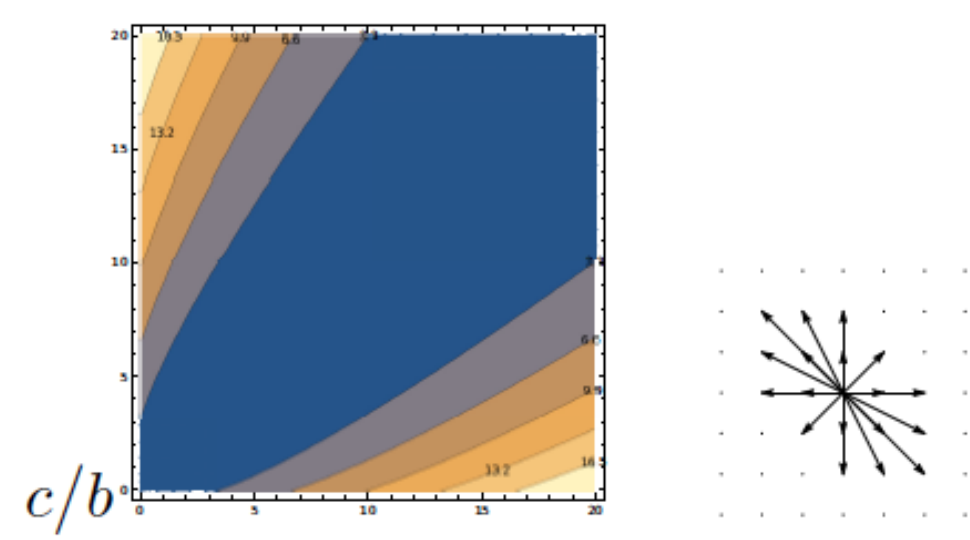


Fig. 1. Contour plots of the transmission disequilibrium test statistic for trio families and the possible moves of (b, c) .

Differential Privacy

Definition 1. (ϵ -Differential Privacy [1])

A randomized mechanism M is ϵ -differentially private if, for all datasets D and D' , which differ in only one family and any $S \subseteq \text{range}(M)$,

$$\Pr[M(D) \in S] \leq e^\epsilon \cdot \Pr[M(D') \in S].$$

Definition 2.

(Sensitivity for the Exponential Mechanism [2])

Let $\mathcal{D}^M$ be the collection of all datasets with M SNPs; then, the sensitivity of a score function $u : \mathcal{D}^M \times \{1, 2, \dots, M\} \rightarrow \mathbb{R}$ is

$$\Delta u = \max_r \max_{D, D'} |u(D, r) - u(D', r)|,$$

where $r \in \{1, 2, \dots, M\}$ and $D, D' \in \mathcal{D}^M$ differ in a single family.

The Shortest Hamming Score

Definition 3. (The SHD score)

Given a predefined threshold $c^* > 0$, the SHD score for i -th data D_i ($i = 1, 2, \dots, M$) is

$$d_{\text{SH}}(D_i, i) = \begin{cases} 0, & \text{if } T_i \geq c^* \text{ and } \exists D'_i, T'_i < c^*, \\ 1 + \min d_{\text{SH}}(D'_i, i), & \text{if } T_i \geq c^* \text{ and } \nexists D'_i, T'_i < c^*, \\ -1 + \max d_{\text{SH}}(D'_i, i), & \text{if } T_i < c^*, \end{cases}$$

where T_i and T'_i are the test statistics obtained from D_i and D'_i , respectively, and $D_i, D'_i \in \mathcal{D}^M$ differ in a single family. For $i \notin \{1, \dots, M\}$, $d_{\text{SH}}(D_i, i) = -\infty$.

Exact Algorithm

Algorithm 1 Exact algorithm to find the SHD score for TDT statistics.

Input: Information about a single SNP, that is, $n_1, n_2, n_3, n_4, n_5, n_6$, and the threshold c^* for the TDT statistics.

Output: The SHD score in one SNP.

```

1:  $T = (n_1 - n_2 + 2n_4 - 2n_5)^2 / (n_1 + n_2 + 2n_3 + 2n_4 + 2n_5)$ 
2:
3: if  $T < c^*$  then
4:   Increase the number of families with  $(b, c) = (2, 0)$ .
5:    $d_1 = 0$ ,  $N_k = n_k (k = 1, \dots, 6)$ 
6:   while  $T < c^*$  do
7:     Check the value of  $N_5, N_2, N_3, N_6$ , and  $N_1$  in that order, and if a value greater than 0 is found, decrease it by one and continue to the next step.
8:      $N_4 \leftarrow N_4 + 1$ 
9:      $T = (N_1 - N_2 + 2N_4 - 2N_5)^2 / (N_1 + N_2 + 2N_3 + 2N_4 + 2N_5)$ 
10:     $d_1 \leftarrow d_1 - 1$ 
11:   end while
12:
13:   Increase the number of families with  $(b, c) = (0, 2)$ .
14:    $d_2 = 0$ ,  $N_k = n_k (k = 1, \dots, 6)$ 
15:   As in the above case, check  $N_4, N_1, N_3, N_6$ , and  $N_2$  in that order, and increase  $N_5$ , then decrease  $d_2$  until  $T \geq c^*$ .
16:
17:   The SHD score is  $\max\{d_1, d_2\}$ .
18:
19: else if  $T \geq c^*$  then
20:   if  $n_1 + 2n_4 > n_2 + 2n_5$  then
21:     As in the case of  $T < c^*$ , check  $n_4, n_1, n_6, n_3$ , and  $n_2$  in that order, and increase  $n_5$  until  $T < c^*$ .
22:   else
23:     Check  $n_5, n_2, n_6, n_3$ , and  $n_1$  in that order, and increase  $n_4$  until  $T < c^*$ .
24:   end if
25:   The SHD score is (the number of steps)  $- 1$ .
26: end if

```

Time Complexity
 $= \mathcal{O}(n)$

Approximation Algorithm

Algorithm 3 Approximation algorithm to find the SHD Score for TDT statistics.

Input: Information about a single SNP, that is, $n_1, n_2, n_3, n_4, n_5, n_6$, and the threshold c^* for the TDT statistics.

Output: The SHD score in one SNP.

```

1:  $b = n_1 + n_3 + 2n_4$ ,  $c = n_2 + n_5 + 2n_5$ 
2:  $T = (b - c)^2 / (b + c)$ 
3: if  $T < c^*$  then
4:   if  $b + c < c^*$  then
5:     The SHD score is  $-\left\lceil \frac{2c^* - (b+c) - |b-c|}{4} \right\rceil$ .
6:   else if  $b + c \geq c^*$  then
7:     The SHD score is  $-\left\lceil \frac{\sqrt{(b+c)c^*} - |b-c|}{4} \right\rceil$ .
8:   end if
9: else if  $T \geq c^*$  then
10:  The SHD score is  $\left\lceil \frac{|b-c| - \sqrt{(b+c)c^*}}{4} \right\rceil - 1$ .
11: end if

```

Time Complexity
 $= \mathcal{O}(1)$

Run Time

10,000
times faster

Small Cohort
(150 families and 5,000 SNPs)

(I)	exact	appx.
(i)	0.875 s	0.020 s
(ii)	0.972 s	0.019 s

Large Cohort
(5,000 families and 10^6 SNPs)

(II)	exact	appx.
(i)	1081.773 s	4.047 s
(ii)	1338.327 s	4.163 s

Accuracy

Small Cohort

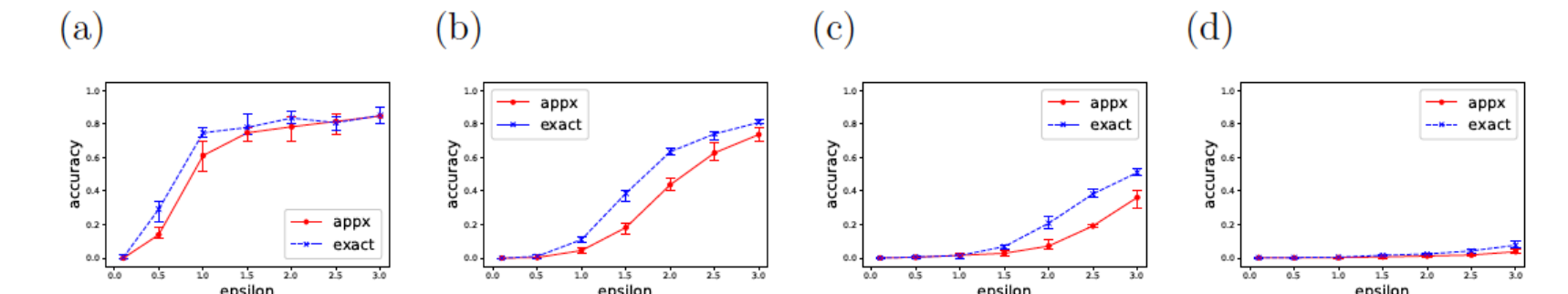


Fig. 2. Accuracy of the top K significant SNPs when (a) $K = 1$, (b) $K = 3$, (c) $K = 5$, and (d) $K = 10$ in case (I)(i).

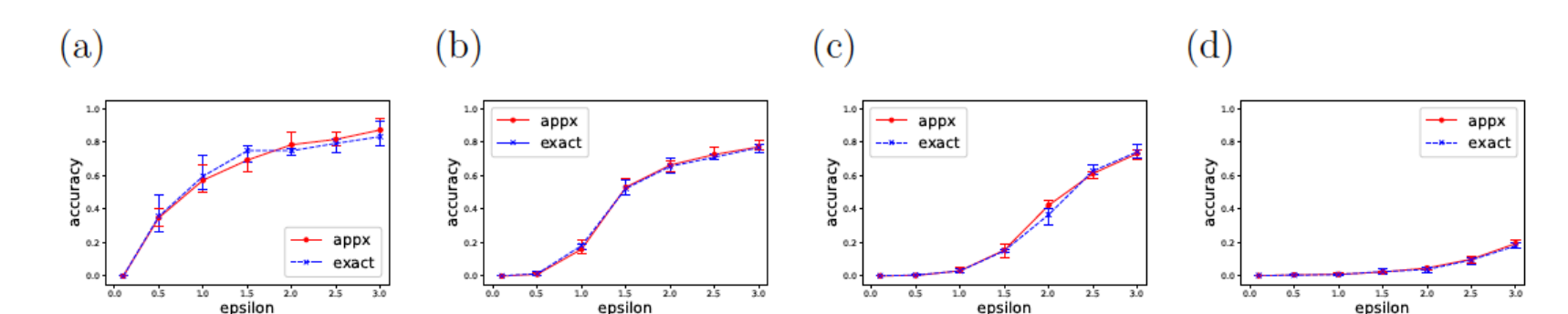


Fig. 3. Accuracy of the top K significant SNPs when (a) $K = 1$, (b) $K = 3$, (c) $K = 5$, and (d) $K = 10$ in case (I)(ii).

Large Cohort

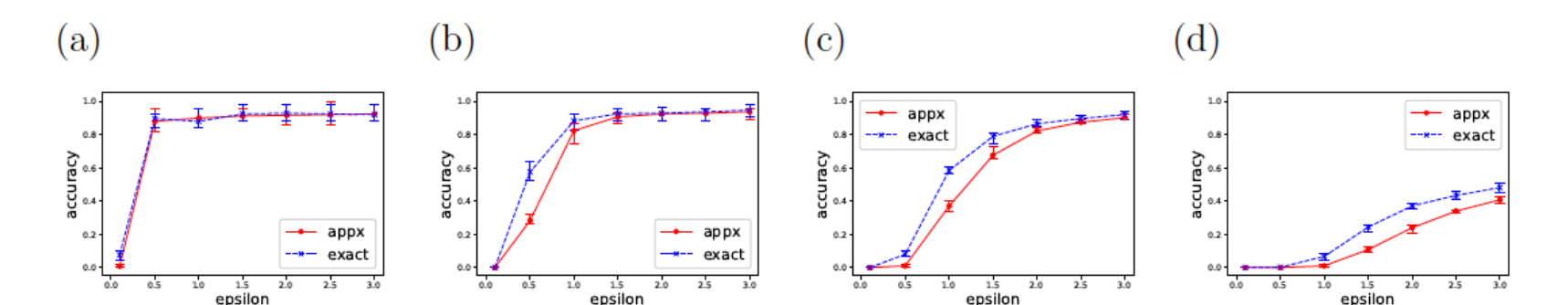


Fig. 4. Accuracy of the top K significant SNPs when (a) $K = 1$, (b) $K = 3$, (c) $K = 5$, and (d) $K = 10$ in case (II)(i).

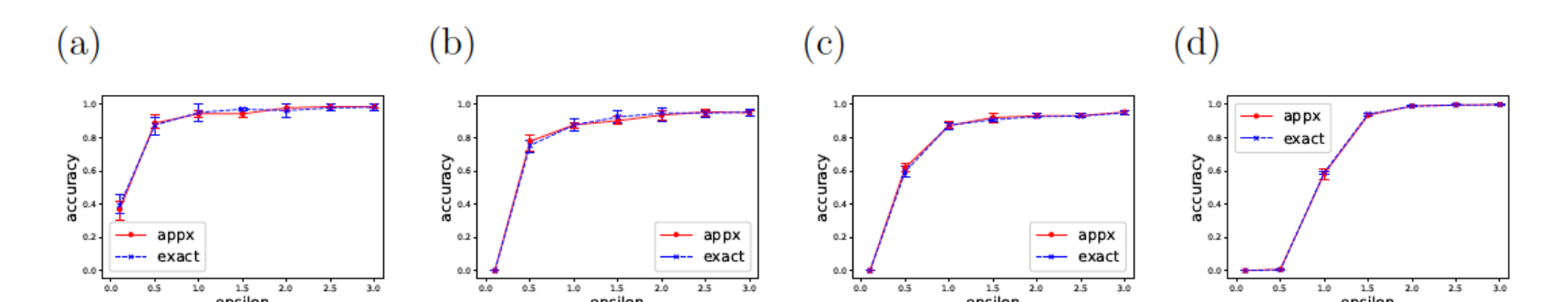


Fig. 5. Accuracy of the top K significant SNPs when (a) $K = 1$, (b) $K = 3$, (c) $K = 5$, and (d) $K = 10$ in case (II)(ii).

Conclusion

- *Sensitivity* of the SHD score obtained by our approximation algorithm is 1.
- Our exact algorithm is about 10,000 times faster than the previous method [3] for a small cohort.
- Our algorithms are the first in the world to be practical even for large cohorts.

Acknowledgement

This work was supported by JSPS KAKENHI Grant 20H05967, 20K21827, 21H05052. and JST CREST Grant JPMJCR1402JST.

References

- [1] Cynthia Dwork. Differential privacy. *Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, (eds) Automata, Languages and Programming*, 4052, 2006.
- [2] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94–103, Providence, RI, USA, October 2007.
- [3] Meng Wang, Zhanglong Ji, Shuang Wang, Jihoon Kim, Hai Yang, Xiaoqian Jiang, and Lucila Ohno-Machado. Mechanisms to protect the privacy of families when using the transmission disequilibrium test in genome-wide association studies. *Bioinformatics*, 33(23):3716–3725, 2017.
- [4] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiayong Zhou. Learning your identity and disease from research papers: Information leaks in genome wide association study. In *CCS'09*, pages 534–544, Chicago, Illinois, USA, November 2009.

Contact information:
Akito Yamamoto
a-ymmt@ims.u-tokyo.ac.jp