Privacy-Enhancing Technologies for Disease Risk Tests Based on Genomic and Non-Genomic Data

1. Motivations

- Genomic data provides opportunities for substantial improvements in diagnosis and preventive medicine.
- Individual's predisposition to disease depends on genomic variations.
- Non-genomic attributes of individuals also contribute significantly to their disease risks.

PRIVACY THREATS DUE TO GENOMIC INFORMATION LEAKAGE:

- Revelation of predisposition to diseases, ethnicity, paternity, filiation, etc.
- Genetic discrimination.
- Denial of access to health insurance, mortgage, education, and employment.

GOALS:

- Protect the privacy of patients' genomic data and non-genomic data on a centralized bio-bank.
- Allow different health stakeholders to access only to the medical data they need (or they are authorized for).
- Allow different health stakeholders to perform some computations on the encrypted data in a privacy-preserving fashion in a reasonable time.

2. Genomic Background

The human genome has approximately 3 billion letters.

- Single Nucleotide Polymorphisms (SNPs): DNA variations, occurring when a single nucleotide differs between members of the same species.
- Potential nucleotides for a SNP position are called alleles.
- A disease risk test is done by analyzing particular SNPs along with other non-genomic risk factors.
- Each SNP contributes to the disease risk in a different amount.
- 40 million approved SNPs in the human population.
- Each patient carries around 4 million SNPs out of 40 million – real SNPs of the patient.
- 75 real SNPs enable the attacker to identify a person.

3. Proposed Framework

4. Disease risk test

A. GENETIC RISK COMPUTATION THROUGH A PRIVATE LOGISTIC REGRESSION MODEL

Patient (P)’s SNPs:

\[
\begin{align*}
\text{Markers for disease X: } & S_{NP_1}, \ldots, S_{NP_2}, \ldots, S_{NP_M} \\
\text{Contributions of markers: } & \beta_1, \beta_2, \ldots, \beta_M \\
\text{P's genetic risk for disease X: } & [S] = \alpha + \sum_{i=1}^{M} \beta_i S_{NP_i}(X)
\end{align*}
\]

B. GENETIC RISK CATEGORIZATION

- For clinical use explanatory variables like the genetic risk should be categorized based on their risk group.
- A private preserving comparison algorithm between SPU and MU allows to compare encrypted values.

C. FINAL DISEASE RISK COMPUTATION

- Let \([G(S,b)]\) be the encrypted result of the comparison between \(S\) and \(b\); thus the encrypted genetic regression coefficient \([\beta_i]\) can be computed as follows:

\[
[\beta_i] = \beta_i(1 - G(S,b_i)) + \sum_{j=1}^{K} \beta_j G(S,b_{j-1}) + \beta_k G(S,b_{K-1})
\]

5. Evaluation on Real Data

- Intel Core i7-2620M CPU with 2.70 GHz processor.
- Size of the security parameter: 4096 bits.
- Real SNP profiles from 1000 Genomes Project.
- Coronary artery disease (CAD) risk factors (23 SNPs, 14 non-genomic factors).
- Java implementation.

<table>
<thead>
<tr>
<th>Complexity of the Proposed System</th>
<th>Encryption</th>
<th>Storage</th>
<th>Computation of disease risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>380 ms./attribute (with pre-compute values: 5.168 ms. attribute)</td>
<td>31.2 GB per patient</td>
<td>230 sec (23 SNPs)</td>
<td>Total: 373.432 sec</td>
</tr>
<tr>
<td>3.390 sec (3 comparisons)</td>
<td>Privacy-preserving integer comparison</td>
<td>30 sec (14 environmental factors)</td>
<td></td>
</tr>
</tbody>
</table>

Appendix A: GOALS:

- Information leakage.
- Privacy threats due to genomic data.
- Information leakage in the genomic PSG.
- Privacy threats due to non-genomic data.
- Information leakage in the non-genomic PSG.
- Privacy threats due to information leakage in the individual patient.
- Privacy threats due to information leakage in the complex system.

Appendix B: SKETCH:

- SKETCH 1: PERSONALIZATION OF GENOMIC RISK.
- SKETCH 2: GENOMIC RISK PRESERVING CATEGORIZATION.
- SKETCH 3: ENCRIPTED GENOMIC RISK.
- SKETCH 4: INFORMATION LEAKAGE IN THE GENOMIC PSG.
- SKETCH 5: INFORMATION LEAKAGE IN THE NON-GENOMIC PSG.
- SKETCH 6: INFORMATION LEAKAGE IN THE INDIVIDUAL PATIENT.
- SKETCH 7: INFORMATION LEAKAGE IN THE COMPLEX SYSTEM.

Appendix C: METRICS:

- Data Privacy Metrics.
- Information Disclosure Metrics.
- Security Metrics.
- Privacy-Preserving Metrics.
- Performance Metrics.

Appendix D: TECHNICAL REPORTS:


Appendix E: ACKNOWLEDGEMENTS:

- Acknowledgments for funding and support.
- Acknowledgments for collaborations and partnerships.
- Acknowledgments for contributions from colleagues and partners.
- Acknowledgments for contributions from institutions and organizations.
- Acknowledgments for contributions from individuals and entities.

Appendix F: REFERENCES:

- Reference 1: Privacy-Preserving Computation of Disease Risk by Using Genomic and Non-Genomic Data.
- Reference 5: Privacy-Preserving Computation of Disease Risk by Using Genomic and Non-Genomic Data.

Appendix G: APPENDIX:

- Appendix A: Privacy-Preserving Computation of Disease Risk by Using Genomic and Non-Genomic Data.
- Appendix B: Privacy-Preserving Computation of Disease Risk by Using Genomic and Non-Genomic Data.
- Appendix C: Privacy-Preserving Computation of Disease Risk by Using Genomic and Non-Genomic Data.
- Appendix D: Privacy-Preserving Computation of Disease Risk by Using Genomic and Non-Genomic Data.
- Appendix E: Privacy-Preserving Computation of Disease Risk by Using Genomic and Non-Genomic Data.
- Appendix F: Privacy-Preserving Computation of Disease Risk by Using Genomic and Non-Genomic Data.
- Appendix G: Privacy-Preserving Computation of Disease Risk by Using Genomic and Non-Genomic Data.