Identifying and Addressing Privacy Challenges of Health Data

Thesis summary – Yamane EL ZEIN

HEC Lausanne/01 – 2025

Health-related data (HRD) is generated in the clinical setting through the use of healthcare information systems by practitioners, and in the personal setting, through the use of digital services. Although HRD has many benefits in the advancement of healthcare, they are considered highly sensitive. This thesis explores and addresses multiple aspects of HRD privacy.

First, we design, implement, and evaluate an efficient and privacy-preserving protocol for collaborative training of decision tree models on distributed biomedical datasets. Our protocol achieves a good balance between model accuracy, efficiency, and privacy, making it suitable for use in the healthcare setting.

Second, we examine *shadow* HRD: HRD generated and/or processed by individuals by using general-purpose digital tools outside of a professional healthcare information system. These can be overlooked when enforcing data protection laws pertaining to HRD, as they are difficult to detect. We identify and categorize a broad variety of user behaviors that lead to the creation of *shadow* HRD. We also analyze the privacy policies of popular service providers, and show that although some of them do treat *shadow* HRD as sensitive data, many do not even acknowledge their collection through their service.

Third, we assess the prevalence of *shadow* HRD-creating behaviors, users' awareness and concerns with respect to *shadow* HRD, and the relevant privacy-utility trade-offs. We find that users deem that the benefits of using general-purpose tools for managing their health outweigh the privacy risks to their health data. Furthermore, users rarely resort to privacy protection measures when doing so. This highlights the need for better protections at the service providers' and legal levels.

Les données relatives à la santé (DRS) sont générées dans le cadre clinique, par l'utilisation par les praticien.ne.s de systèmes d'information pour les soins de santé, et dans le cadre personnel, par le biais de services numériques. Bien que les DRS présentent de nombreux avantages pour l'amélioration des soins de santé, elles sont très sensibles. Cette thèse aborde multiples aspects de la confidentialité des DRS.

Premièrement, nous concevons, implémentons, et évaluons un protocole efficace et assurant la confidentialité, pour l'apprentissage collaboratif de modèles d'arbres de décision sur des données biomédicales réparties. Ce protocole présente un équilibre entre précision du modèle, efficacité du calcul, et confidentialité, ce qui le rend adapté à une utilisation dans le domaine de santé.

Deuxièmement, nous examinons les DRS fantômes : DRS générées et/ou traitées à l'aide d'outils numériques à usage général, en dehors d'un système d'information professionnel de santé. Étant difficiles à détecter, ces données peuvent être négligées lors de l'application des lois sur la protection des données. Nous identifions et catégorisons les comportements d'utilisat.eur.rice.s conduisant à la création de DRS fantômes. Nous analysons ensuite les politiques de confidentialité de fournisseurs de services, et montrons que bien que certains traitent les DRS fantômes comme sensibles, la plupart ne reconnaissent même pas leur collecte.

Troisièmement, nous évaluons la prévalence des comportements créateurs de DRS fantômes, la sensibilisation et les préoccupations des utilisat.eur.rice.s quant à ces données, et les compromis entre vie privée et utilité. Nous constatons que les utilisat.eur.rice.s considèrent que les avantages de l'utilisation d'outils à usage général pour gérer leur santé l'emportent sur les risques d'atteinte à la vie privée. En outre, iels prennent rarement des mesures de protection lorsqu'iels adoptent de tels comportements. Cela souligne la nécessité d'une meilleure protection au niveau des fournisseurs de services et au niveau juridique.

UNIL | Université de Lausanne
HEC Lausanne