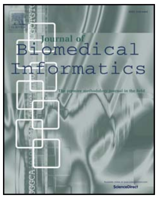
Journal of Biomedical Informatics 79 (2018) 1–6

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Special Communication

Are privacy-enhancing technologies for genomic data ready for the clinic? A survey of medical experts of the Swiss HIV Cohort Study

Jean-Louis Raisaro[a], Paul J. McLaren[b,c], Jacques Fellay[d,e], Matthias Cavassini[f], Catherine Klersy[g], Jean-Pierre Hubaux[a,*], the Swiss HIV Cohort Study

[a] School of Computer Communications Sciences, École Polytechnique Fédérale de Lausanne, Switzerland
[b] J.C. Wilt Infectious Diseases Research Centre, National Microbiology Laboratories, Public Health Agency of Canada, Winnipeg, Canada
[c] Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, Canada
[d] School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland
[e] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[f] Division of Infectious Diseases, Lausanne University Hospital, Switzerland
[g] Service of Biometry and Clinical Epidemiology, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

ARTICLE INFO

Keywords:
Genomic privacy
Clinical genomics
Genetic testing
Encryption
Genetic test reporting
Privacy-enhancing technologies
Homomorphic encryption
Survey

ABSTRACT

Purpose: Protecting patient privacy is a major obstacle for the implementation of genomic-based medicine. Emerging privacy-enhancing technologies can become key enablers for managing sensitive genetic data. We studied physicians' attitude toward this kind of technology in order to derive insights that might foster their future adoption for clinical care.

Methods: We conducted a questionnaire-based survey among 55 physicians of the Swiss HIV Cohort Study who tested the first implementation of a privacy-preserving model for delivering genomic test results. We evaluated their feedback on three different aspects of our model: clinical utility, ability to address privacy concerns and system usability.

Results: 38/55 (69%) physicians participated in the study. Two thirds of them acknowledged genetic privacy as a key aspect that needs to be protected to help building patient trust and deploy new-generation medical information systems. All of them successfully used the tool for evaluating their patients' pharmacogenomics risk and 90% were happy with the user experience and the efficiency of the tool. Only 8% of physicians were unsatisfied with the level of information and wanted to have access to the patient's actual DNA sequence.

Conclusion: This survey, although limited in size, represents the *first* evaluation of privacy-preserving models for genomic-based medicine. It has allowed us to derive unique insights that will improve the design of these new systems in the future. In particular, we have observed that a clinical information system that uses homomorphic encryption to provide clinicians with risk information based on sensitive genetic test results can offer information that clinicians feel sufficient for their needs and appropriately respectful of patients' privacy.

The ability of this kind of systems to ensure strong security and privacy guarantees and to provide some analytics on encrypted data has been assessed as a key enabler for the management of sensitive medical information in the near future. Providing clinically relevant information to physicians while protecting patients' privacy in order to comply with regulations is crucial for the widespread use of these new technologies.

1. Introduction

Data breaches in healthcare are increasingly costly and frequent, and patients, care givers and hospital administrations are suffering their devastating effects [1,2]. Most of these breaches are due to criminal cyber attacks and internal threats that target healthcare organizations and exploit glitches of their IT security systems or human errors [3].

In this complicated environment, guaranteeing the security and privacy of electronic health records (EHRs) is becoming a major challenge for healthcare providers trying to avoid the erosion of their patients' trust. The privacy impact of these attacks will only increase when EHRs are linked to genetic information needed for precision medicine. Unlike many other medical data, genome sequences are unique and permanent: they cannot be de-identified or replaced when

[*] Corresponding author at: School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Station 14, EPFL IC ISC LCA1, BC 207, 1015 Lausanne, Switzerland.

E-mail address: jean-pierre.hubaux@epfl.ch (J.-P. Hubaux).

https://doi.org/10.1016/j.jbi.2017.12.013
Received 12 June 2017; Received in revised form 21 December 2017; Accepted 23 December 2017
Available online 10 January 2018

compromised [4–6].

Yet, despite the awareness of this growing threat to patient data, most healthcare organizations continue to depend mainly upon outdated policies and procedures that are insufficient to address and minimize these attacks [7,8]. Healthcare providers struggle to adopt modern privacy-enhancing technologies (PETs) that are already well-established in the IT security and privacy community [9].

Limited investments in technical security measures partly explain this slow adoption, but the cultural gap between the medical and IT security and privacy communities also represents a significant roadblock for the deployment of PETs in the clinic.

In 2016, we began to bridge this gap by proposing what was possibly the first practical clinical deployment of a privacy-preserving model for genomic testing [10]. Our study showed the applicability, in a real operational setting, of sophisticated privacy-preserving techniques based on homomorphic encryption for genetic testing with ancestry inference and delivery of interpreted information to clinicians. In particular, we demonstrated the computation of various genetic tests, such as prediction of abacavir hypersensitivity and cardiovascular risk, directly on the encrypted genotype data from 230 HIV-positive individuals. Homomorphic encryption is the state-of-the-art technology that enables computation on encrypted data without exposing the clear data decrypting it first [11,12]. Only individuals who possess the decryption key can access the final result of the computation. However, a barrier to wide deployment of this technique in the medical setting is its acceptance and ongoing usage by clinicians who are concerned that information systems can provide useful and actionable information while respecting privacy.

In this survey, we present the insights collected from HIV specialists from the outpatient clinics of the Swiss HIV Cohort Study (SHCS) [13] that have been testing our system for one year. Building on our previous work, we describe end-users' perception of genome privacy and their attitude toward the use of sophisticated PETs in an operational clinical setting. In particular, we aimed at identifying the key aspects of our model that could represent general drivers for a faster adoption of privacy-enhancing solutions for genomic data protection.

2. Materials and methods

2.1. Privacy-preserving model for genomic testing

We begin by describing the privacy-preserving model for genomic testing that we proposed in [10] and that is the subject of this survey.

Our model consists of four main parties: (i) the patients; (ii) a certified institution (CI) responsible for genotyping, management of cryptographic keys, and encryption of patients' genetic data; (iii) a central data center (DC) where the encrypted variants are stored; and (iv) a set of outpatient clinics wishing to perform genetic tests on the patients. We consider the certified institution to be trusted, as it has access to the unprotected raw genetic variants as output of the sequencing process, and both the DC and the clinics to be honest-but-curious. This means that the DC and the clinics honestly follow the genomic testing protocol but they can be compromised by adversaries (e.g., insiders or hackers) that might passively infer sensitive information about the patients. We assume also that the clinics do not collude with the DC.

The privacy-preserving genomic testing protocol makes use of additively homomorphic encryption, deterministic encryption, proxy re-encryption, and secure two-party protocols in order to provide end-to-end protection of patients' data confidentiality and thwart the attacks of honest-but-curious adversaries. The protocol consists of an "offline" phase and an "online" phase. During the offline phase, the CI generates, for each patient, a pair of asymmetric cryptographic keys (public and private) for a modified version of the Paillier homomorphic encryption scheme [14], and individually encrypts each patient's variant calls. The Paillier homomorphic encryption scheme is a probabilistic encryption scheme that provides semantic security (also known as the indistinguishability of ciphertexts) based on the quadratic residuosity assumption. We used a 4096-bit keys as recommended by the US National Institute of Standards and Technologies in order to provide security for the next 30-plus years. The encrypted genetic variants are stored at the DC and the private key of each patient is randomly split into two shares, with one share assigned to the DC and the other to the clinics. Such a key-splitting technique is needed in order to prevent the confidentiality of the genetic data stored at the DC from being compromised if either the DC or one of the clinics is compromised. In other words, splitting the private key among multiple parties ensures trust decentralization as no single entity in the system needs to be trusted, thus avoiding a single point of failure. Conversely, the public key is sent in full to each party. At the end of the offline phase, the CI deletes the raw genetic data and all the cryptographic keys. In the online phase, any clinic wishing to perform a genomic test on a given patient starts a privacy-preserving protocol with the DC. During such protocol, the clinic uses homomorphic encryption to securely combine the encrypted variant calls stored at the DC and compute the patient's encrypted genetic risk score, without ever decrypting the data.

Computation under encryption is only possible thanks to the homomorphic properties of the Paillier cryptosystem. We note that with any standard non-homomorphic cryptosystem, the clinic would have to decrypt each variant call before combining them in the clear in order to obtain the patient's genetic risk score, thus exposing the patient's genomic variants. Instead, with this secure testing protocol, at any point in time neither party can see the patient's genetic variants in the clear. Once the encrypted genetic risk score is computed, the DC uses its share of the patient's private key to partially decrypt the result and send the resulting ciphertext to the clinic. Finally, the clinic uses the other share of the patient's secret key to fully decrypt the result.

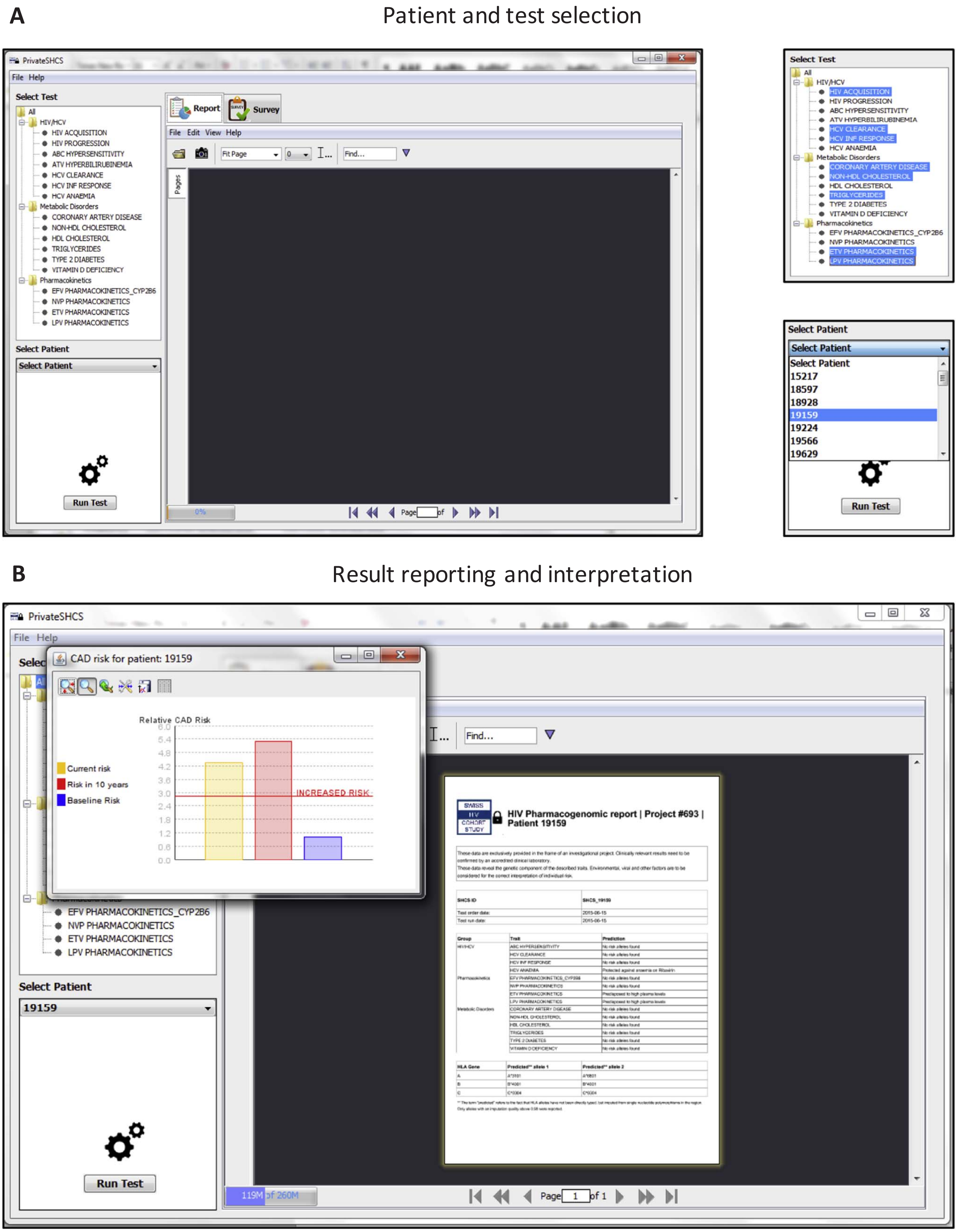
In the setting of the Swiss HIV Cohort Study (SHCS), we considered genetic tests targeting 71 markers that were informative for 17 traits relevant to HIV outcomes. Clinically informative markers fell into three categories: (i) HIV/hepatitis C virus progression and response to therapy; (ii) pharmacokinetics of efavirenz, nevirapine, etravirine or lopinavir; and (iii) metabolic traits including vitamin D deficiency, coronary artery disease, cholesterol and triglyceride levels, and type 2 diabetes. Testing included single and multimarker deterministic tests and multimarker risk scores. Test results, securely computed under encryption by using homomorphic techniques, were interpreted and provided to clinicians so that when significant genetic markers or risk scores were observed, an alert specific to the test was returned.

For some tests, however, there is strong correlation between the result and the underlying causal genotype (e.g., abacavir hypersensitivity). The inclusion of a large number of these tests could be risky for patients' privacy as the knowledge of 70–80 variants is enough to uniquely identify an individual [15]. Yet, this was not the case for the proposed system as the results of only six tests (HIV acquisition, abacavir hypersensitivity, antiretroviral hyperbilirubinemia, HCV clearance, HCV interferon response and HCV anaemia) were revealing the underlying variants. In the case where the results of many such tests are to be included in the same reports, other techniques, such as result obfuscation, can be used on top of this model.

2.2. Setting and participants

This survey was approved by the Scientific Board of the SHCS and by the affiliated outpatient clinics. It took place in five principal hospitals in Switzerland: three university hospitals (Zurich, Lausanne and Basel) and two cantonal hospitals (St. Gallen and Lugano). All physicians who were members of the SHCS were invited to participate in the survey. To represent a "real life" situation, no specific training about PETs or IT security was organized for the participants prior to the study.

All participants had access to the system for privacy-preserving genetic testing through a Java front-end application installed at each participating hospital (see Fig. 1). The application comes with a



SWISS HIV COHORT STUDY

HIV Pharmacogenomic report | Project #693 | Patient 19159

These data are exclusively provided in the frame of an investigational project. Clinically relevant results need to be confirmed by an accredited clinical laboratory.
These data reveal the genetic component of the described traits. Environmental, viral and other factors are to be considered for the correct interpretation of individual risk.

SHCS ID	SHCS_19159
Test order date:	2015-06-15
Test run date:	2015-06-15

Group	Trait	Prediction
HIV/HCV	ABC HYPERSENSITIVITY	No risk alleles found
	HCV CLEARANCE	No risk alleles found
	HCV INF RESPONSE	No risk alleles found
	HCV ANAEMIA	Protected against anaemia on Ribavirin
Pharmacokinetics	EFV PHARMACOKINETICS_CYP2B6	No risk alleles found
	NVP PHARMACOKINETICS	No risk alleles found
	ETV PHARMACOKINETICS	Predisposed to high plasma levels
	LPV PHARMACOKINETICS	Predisposed to high plasma levels
Metabolic Disorders	CORONARY ARTERY DISEASE	No risk alleles found
	NON-HDL CHOLESTEROL	No risk alleles found
	HDL CHOLESTEROL	No risk alleles found
	TRIGLYCERIDES	No risk alleles found
	TYPE 2 DIABETES	No risk alleles found
	VITAMIN D DEFICIENCY	No risk alleles found

HLA Gene	Predicted** allele 1	Predicted** allele 2
A	A*3101	A*6801
B	B*4001	B*4001
C	C*0304	C*0304

** The term "predicted" refers to the fact that HLA alleles have not been directly typed, but imputed from single nucleotide polymorphisms in the region. Only alleles with an imputation quality above 0.98 were reported.

Fig. 1. Front-end graphical user interface. (A) Through this graphical user interface the physician can select one or multiple tests to be run on a given patient. (B) Encrypted results are decrypted, interpreted and presented as a standardized text report.

graphical user interface enabling physicians to select patients using a unique identifier and the genetic tests to be performed on the patient's encrypted genotype (see Fig. 1A). The request is sent to the system back-end, deployed at the SHCS Data Center in Lausanne where patients' data are stored encrypted in a relational database, and securely processed using the homomorphic properties of the underlying cryptosystem. Patients' genomic data are never decrypted and the encrypted test results are returned to the local front-end for decryption. Finally, a standardized text report indicating an increased or decreased risk of the tested trait due to genetic factors, or a result of "no relevant alleles found" (see Fig. 1B) is presented to the physician.

After using our system, participants were asked to answer an electronic survey directly embedded in the front-end application. Prior to filling the questionnaire, participants had to read a short description of the project illustrating the key elements of the privacy-preserving system they were testing. Upon completion, surveys were automatically sent to the SHCS Data Center for incorporation into an anonymized database.

2.3. Questionnaire

The questionnaire consisted of two sections. The first section included four questions covering participants' basic demographics such as age, gender, grade (options: Resident, Attending and Head of Department) and SHCS center. This information was necessary to verify that our sample was representative of the general physician population.

The second section of the questionnaire aimed at obtaining insights on participants' acceptance level regarding our privacy-preserving system for genetic testing, and more generally, on their attitude toward the adoption of privacy-enhancing technologies for the protection of genetic data in the clinical context. As such, we studied physicians' acceptance level under three different angles represented by the following three observable variables: *clinical utility*, *privacy concerns* and *system usability*. For each of these variables we proposed a set of statements, representative of the concept, about which participants had to indicate their level of agreement with a score from 0 to 4, where 0 means "strongly disagree", 1 means "disagree", 2 means "neutral", 3 means "agree" and 4 means "strongly agree". For the three above-mentioned variables we designed four, six and four statements, respectively.

Prior to this study, we pilot-tested the questionnaire with two attending doctors working on infectious diseases at a clinic not involved in our study. The goal was to assess if the statements were fully understandable and without ambiguity. Both doctors completed all sections without reporting any ambiguity. Hence, the questionnaire was used for the study without further validation.

2.4. Quantitative data analysis

We analyzed survey responses with Python Data Analysis Library and with STATA™ software version 14.2. For each statement, we first computed the proportions of physicians per agreement level. Then, for each participant we aggregated his/her scores grouped by variable (i.e., *clinical utility*, *privacy concerns*, *system usability)* in order to obtain a triplet score where each element represented the sum of the scores for the statements related to a given concept. We hypothesize that the three observed variables are the collective expression of the general acceptance level of our system. This is an unmeasured (latent) variable whose relationship with the 3 indicators can be quantified via Confirmatory Factor Analysis (CFA) [16] in the framework of structural equation models (SEM). We derived the relative importance of each indicator from SEM, by estimating standardized coefficients. We assessed the goodness of fit of the model using the following indices: Comparative Fit Index (> 0.95), Tucker Lewis Index (> 0.95), and root mean square error of approximation (< 0.06).

3. Results

3.1. Participants characteristics

Of 55 SHCS-affiliated doctors invited to the study, 38/55 (69%) tested the proposed system at least once and completed the survey. There were 8 (21%) females and 30 (79%) males with mean age of 41 years and a maximum of 65. Of the 38 participants, 12 (31%) were attending physicians, 11 (29%) were heads of department and 15 (40%) were residents. Our sample is marginally younger and with some over-representation of male than the overall physician population as compared to the 2016 Report of the Swiss Doctors' Federation (FMH), which reported that the mean age of physicians is 46 and that 60% of physicians in Switzerland are male [17].

3.2. Quantitative results

We report the distributions of physicians' scores for the different statements in the survey in Fig. 2A. We observed that statements for all three variables received high consensus (i.e., agreement level of 3 or 4) from a large majority of the physicians. Concerning the *clinical utility* of our system, 71% of participants considered the information provided useful and 68% that it was worth a therapy modification when actionable. Only 8% of physicians were unsatisfied with the level of information and wanted to have access to the patient's actual DNA sequence.

Regarding the magnitude of participants' *privacy concerns*, 76% agreed or strongly agreed that the human genome actually contains privacy-sensitive information. Physicians showed similar concerns for potential exposure of their patients' clinical (65%) and genomic (60%) data. Moreover, almost all (92%) thought that it is the responsibility of the healthcare institution storing the genetic data to make sure that only authorized people can access them. Finally, we observed that for a majority of physicians (68%), a better diagnosis and treatment should not be at the expense of privacy and that in general (71%), physicians prefer to obtain only the information necessary for the diagnosis in order to avoid the risk of incidental findings.

Concerning the *system usability*, participants had a user-friendly experience with our system. Physicians had no perception of the sophisticated privacy-preserving techniques running behind the scenes as the encryption, result decryption and cryptographic key management were performed automatically by the system. Only 10% of those surveyed required technical support and were skeptical about the use of similar privacy-preserving systems in their daily clinical practice mainly because of the reduced amount of provided information (disagreement with the principle of "data minimization") and the potentially involved interaction with these tools.

Fig. 2B shows the results of CFA. The reported standardized weights measure the relative importance of each of *clinical utility*, *privacy concerns* and *system usability* for participants' overall acceptance of the system. We can observe that the impact on the acceptance level of *clinical utility* (0.8) is twice the impact of *privacy concerns* (0.39).

4. Discussion

Through this survey we were able to obtain unprecedented insights about the attitude of real end-users toward a privacy-preserving system for genetic testing and to derive a more general understanding about the key requirements for the adoption of privacy-enhancing technologies in a routine clinical context.

The first important take-home message is that new systems based on sophisticated privacy-enhancing technologies, such as homomorphic encryption, that can perform analytics on genomic data and provide results without revealing sensitive information have the potential to be useful. Despite their inherent complexity, tools as the one evaluated in

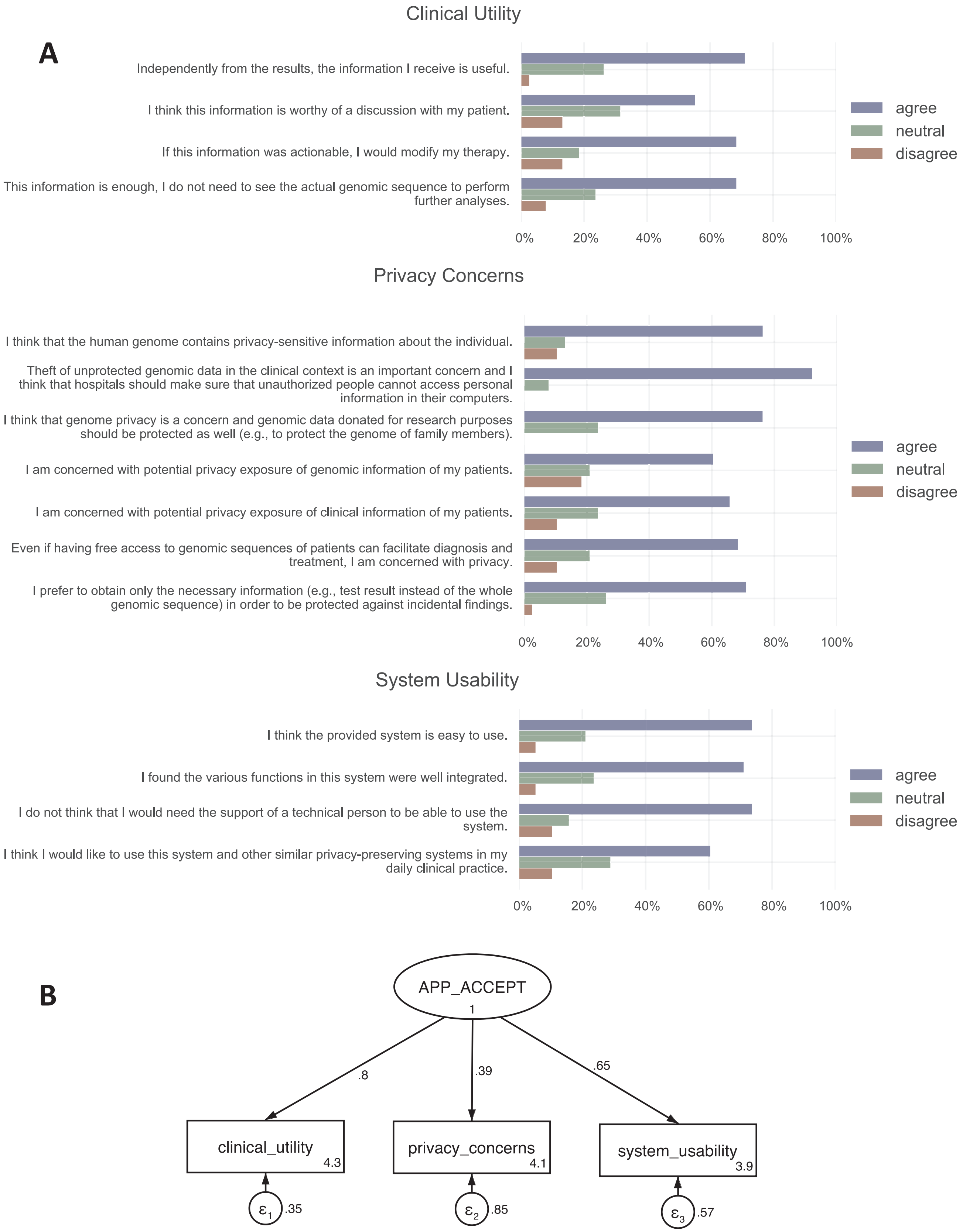


Fig. 2. Quantitative Results. (A) Distributions of scores per claim grouped by observed variable. The category "agree" includes scores > 2, while the category "disagree" includes scores < 2. (B) Confirmatory factor model where the oval indicates the latent (unmeasured) overall acceptance of the system or factor, (APP_ACCEPT), rectangles indicate the observed variables and single-headed arrows show causality from the factor to the indicators. Numbers along the arrows are the SEM standardized weights that indicate the relative importance of the observed variables as measures of general acceptance of the system. Numbers at the epsilon circles are the error term coefficients of the SEM. Goodness of fit criteria of CFA were satisfied.

this study are seen by a large majority of the survey participants as important enablers for the implementation of genomic-based medicine.

As expected, our results show that more than two thirds of the survey participants acknowledged that protecting patient's genetic privacy is crucial when dealing with genetic data both in a clinical and research environment. Privacy-preserving system such as ours have the potential to improve the management of patients' sensitive data, not only because they provide strong privacy and security guarantees, but also because they can protect physicians from undesirable liability issues (e.g., in the case of incidental findings not reported to patients) by limiting their access only to the necessary information.

Our confirmatory factor analysis indicates that, in addition to the ability of ensuring privacy and security of genetic data, the key requirement for the success and deployment of this new kind of medical information systems resides in the ability to make them usable and understandable by end-users and, most importantly, in the clinical value of the information made accessible to physicians. Our study shows that, despite the fundamental tension between data privacy and data access, an acceptable trade-off can often be found. Indeed, providing only the necessary information according the principle of data minimization, as described in the international good practices for genomic data management and in the data protection laws of Switzerland and Europe [18–20], is largely accepted by physicians for this kind of medical applications (e.g., genetic testing) if such information can be clinically useful and actionable.

5. Conclusion

Privacy protection of sensitive medical data and particularly genetic data is an important concern as the healthcare industry is increasingly suffering from data breaches.

In this survey, we have studied for the *first* time physicians' attitude toward the adoption of new privacy-preserving models for using genetic data in the clinic, by evaluating the feedback of 38 HIV specialists on the *first* real-life deployment of a privacy-preserving system for genetic testing based on homomorphic encryption. We have derived unique insights that will guide the design and facilitate the adoption of these systems in the future, by better addressing physicians' requirements. In particular, we have seen that, despite the general skepticisms around the apparent unpracticality and difficulty of these technologies, systems based on cutting-edge privacy-enhancing can be made efficient, deployable and usable in real operation settings. As physicians' primary scope is patients' health, the inherent complexity of these tools should be concealed as much as possible in order to facilitate physicians' work and avoid undesirable overheads.

Finally, we believe that this survey, although limited in size, represents a first step toward the full awareness among physicians, policy makers, hospital administrators and the general public about the existence of sophisticated PETs that can play a significant role in mitigating the incidence of data breaches without preventing the use of the data.

Acknowledgements

The authors thank Yannick Vallet, Franck Youssef, and Bruno Ledergerber for their technical support and Enos Bernasconi, Patrick Schmid, Stoeckle Marcel and Helen Kovari for their help with involving HIV specialists in SHCS outpatient clinics to participate in our survey.

Conflict of interest

The authors declare no conflict of interest.

References

[1] U.S. Department of Health and Human Services Office for Civil Rights, Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected Health Information [Internet]. Available from: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf.
[2] G. Bai, J. Jiang, R. Flasher, Hospital risk of data breaches, JAMA Intern. Med. (2017).
[3] Ponemon Institute, Third Annual Benchmark Study on Patient Privacy & Data Security Sponsored by ID Experts, vol. May, 2016, pp. 50.
[4] A. Harmanci, M. Gerstein, Quantification of private information leakage from phenotype-genotype data: linking attacks, Nat. Methods 13 (3) (2016).
[5] M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, J.-P. Hubaux, De-anonymizing genomic databases using phenotypic traits, PETS 2015 2015 (2) (2015) 1–16.
[6] M. Gymrek, A.L. McGuire, D. Golan, E. Halperin, Y. Erlich, Identifying personal genomes by surname inference, Science (80-) 339 (6117) (2013) 321–324.
[7] H. Kathryn, E.J. Topol, The Health Data Conundrum [Internet], The New York Times. Available from: https://www.nytimes.com/2017/01/02/opinion/the-health-data-conundrum.html?_r=1, 2017 (cited 2017 Apr 10).
[8] J. Kulynych, Is Privacy the Price of Precision Medicine? [Internet], Oxford University Press. Available from: https://blog.oup.com/2017/03/privacy-precision-medicine/, 2017 (cited 2017 Apr 10).
[9] Hayden E. Check, Cloud cover protects gene data, Nature 519 (2015) 400–401.
[10] P.J. McLaren, J.L. Raisaro, M. Aouri, M. Rotger, E. Ayday, I. Bartha, et al., Privacy-preserving genomic testing in the clinic: a model using HIV treatment, Genet. Med. 18 (8) (2016) 814–822.
[11] R.L. Rivest, L. Adleman, M.L. Dertouzos, On data banks and privacy homomorphisms, Found. Secur. Comput. (1978) 169–180.
[12] C. Gentry, Fully homomorphic encryption using ideal lattices, in: Proc 41st Annu. ACM Symp. Symp. Theory Comput. STOC 09, vol. 19, September, 2009, pp. 169.
[13] Swiss HIV Cohort Study [Internet]. Available from: http://www.shcs.ch/.
[14] E. Bresson, D. Catalano, D. Pointcheval, A simple public-key cryptosystem with a double trapdoor decryption mechanism and its applications, ASIACRYPT 2003: Advances in Cryptology, Springer-Verlag, 2003, pp. 37–54.
[15] Z. Lin, A.B. Owen, R.B. Altman, Genomic research and human subject privacy, Science (80-) 305 (5681) (2004) 183–183.
[16] R.B. Kline, Principles and Practice of Structural Equation Modeling, second ed., Guilford Publications, 2005, p. 534.
[17] S. Hostettler, E. Kraft, 36 175 médecins en exercice, 98(13) (2017) 394–400.
[18] Federal Act on Data Protection [Internet]. Available from: https://www.admin.ch/opc/en/classified-compilation/19920153/index.html.
[19] The EU General Data Protection Regulation (GDPR).
[20] The Global Alliance for Genomics and Health, GA4GH Privacy and Security Policy, 2015.