


REVIEW

Open Access



Legal aspects of privacy-enhancing technologies in genome-wide association studies and their impact on performance and feasibility

Alissa Brauneck^{1*}, Louisa Schmalhorst^{1†} , Stefan Weiss², Linda Baumbach³, Uwe Völker², David Ellinghaus^{4†}, Jan Baumbach^{5†} and Gabriele Buchholtz^{1†}

[†]Alissa Brauneck and Louisa Schmalhorst shared first authors.

[†]David Ellinghaus, Jan Baumbach and Gabriele Buchholtz shared last authors.

*Correspondence: alissa.brauneck@uni-hamburg.de

¹Hamburg University Faculty of Law, University of Hamburg, Hamburg, Germany

²Interfaculty Institute of Genetics and Functional Genomics, Department of Functional Genomics, University Medicine Greifswald, Greifswald, Germany

³Department of Health Economics and Health Services Research, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

⁴Institute of Clinical Molecular Biology (IKMB), Kiel University and University Medical Center Schleswig-Holstein, Kiel, Germany

⁵Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany

Abstract

Genomic data holds huge potential for medical progress but requires strict safety measures due to its sensitive nature to comply with data protection laws. This conflict is especially pronounced in genome-wide association studies (GWAS) which rely on vast amounts of genomic data to improve medical diagnoses. To ensure both their benefits and sufficient data security, we propose a federated approach in combination with privacy-enhancing technologies utilising the findings from a systematic review on federated learning and legal regulations in general and applying these to GWAS.

Introduction

'Privacy by design' is an international principle of data protection law which stipulates that privacy measures must be built into the technical and organisational processes which handle personal data. This principle has been laid down in laws in different legislations, e.g. the European Union's General Data Protection Regulation (GDPR) [1] or the California Consumer Privacy Act (CCPA) [2]. In particular, genomic data are highly sensitive [3]. For use in biomedical studies such as genome-wide association studies (GWAS), they often must be shared between institutions. Therefore, to achieve privacy compliance, researchers conducting such studies are required to implement privacy by design to achieve data self-determination. Essentially, this means that contractual agreements to respect privacy are not enough, but instead, researchers must reduce the possibility of privacy violations as much as possible, both through technology and appropriate organisational design. Privacy by design aims to institutionalise privacy at all levels, rather than tinkering with individual processes [4]. However, technology is developing rapidly and privacy by design principles, once formulated, are not necessarily sufficient



© The Author(s) 2024, corrected publication 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

to guarantee a satisfactory level of data protection in the long-term [5]. It is therefore not surprising that in practice, compliance with privacy by design, although necessary, is often perceived as a burden [6]. Challenges associated with the processing of genomic data—e.g. how privacy and research on genomic data can be harmonised, how genome-phenome investigations such as GWAS can be conducted without violating the privacy of the people involved and how individual or combined privacy-enhancing technologies (PET) can be used to meet privacy requirements—have repeatedly been the subject of many papers. For example, Berger and Cho [7] described the shift from traditional privacy approaches for sharing genomic data to advanced privacy-enhancing approaches and their challenges under data protection laws. Erlich and Narayanan [8] examined privacy breaches that are relevant to genomic information, e.g. attribute disclosure attacks via DNA (ADAD), which are particularly relevant for GWAS, as they are especially vulnerable to this form of attack, and appropriate risk mitigation strategies; these, however, do not refer to the legal requirement for privacy protection [8]. In their review, Bonomi et al. [9] analysed the privacy challenges associated with emerging applications for genetic testing performed directly by consumers and what techniques can protect privacy in the context of such analyses. Wan et al. [5] studied the regulations in the EU and the USA on the handling of genetic and genomic data and how the legal differences affect the use of such data, but do not provide a concrete analysis of the legal requirements. Shabani and Marelli [10] referred to codes of conduct or professional society guidance, i.e. ‘soft law’, in order to minimise the risks and offer the greatest possible legal protection for the handling of sensitive data such as genomic data and help to meet the requirements of the GDPR. Mitchell et al. [11] also discussed codes of conduct and additional certification mechanisms under Article 42 GDPR, giving a detailed overview of the legal framework under the GDPR and pointing out various difficulties, such as cross-border data transfers, how to deal with data relating to multiple genetic relatives or the right to rectification when genomic data is inaccurate. Other authors focus on the legal perspective: Quinn and Quinn [12] provided a general evaluation of genetic data under the GDPR and in regard to privacy by design, whilst Brauneck et al. [13] assessed federated learning and privacy-enhancing technologies (PETs) as measures to achieve GDPR compliance.

Our article diverges from prior work in that we trace the principle of privacy by design back to its legal basis and identify the requirements that need to be met before applying them specifically to GWAS on diseases and human traits. On this basis, we analyse each step of these studies and discuss the risks for data subjects associated with them as well as the legal downsides and merits of technical solutions before providing concrete advice on how to fulfil the privacy by design requirements of the GDPR. These requirements are enshrined in Article 25 GDPR and designed to safeguard data subjects’ rights, especially the right to informational self-determination. We focus on GWAS, however, the privacy by design concept applies to all types of studies in which genomic data from individuals are exchanged between different research sites for analysis purposes. We consider the same general privacy risk model as Wang et al. [14]. There are several known types of privacy attacks that are relevant to genomic data sharing, such as membership inference attacks [15, 16], attribution inference attacks [17] and reconstruction attacks [18]. Most commonly, attackers have access to the full or partial genomic

sequences of the target and exploit side information, which usually increases the malicious potential of the attack significantly [14]. Our focus, however, is on general privacy risks, without focussing on specific attacks and aims to mitigate the privacy risks associated with the exchange of highly sensitive data through the use of privacy-enhancing techniques. First, we address the international and European background of privacy by design requirements, then demonstrate which challenges arise in research with genomic data, especially in GWAS with regard to GDPR requirements, and finally present recommendations for future GWAS in the form of privacy-enhancing technologies.

Privacy by design and its impact on genome-wide association studies: a primer

GWAS aim to determine the impact of variation in the genome sequence on physical traits by identifying relationships between genetic variants and phenotypes, such as diseases, disease severity or other human traits. As a result, GWAS can both identify genetic risk factors and improve the standard of medical care [19]. The power of GWAS—especially when analysing common diseases and common variants (and with increasing sample sizes also rare diseases and/or low-frequency variants)—can most effectively be harnessed by studying large datasets from multiple centres with a very high number of participants. This requires data sharing amongst internationally distributed consortia [5, 20–24], which poses a number of legal challenges, not all of which are necessarily unique to GWAS, but result from the large number of participants and consequently large amounts of data that are required for performing GWAS. All of these challenges, which we will investigate in the following, can ultimately be traced back to the requirements of privacy by design.

Privacy by design is far from new [4]. There are many international examples of legislation on how privacy by design might be implemented. In the USA, this principle has been enshrined in, amongst others, the CCPA, and in 2012, the U.S. Federal Trade Commission (FTC), a regulator for antitrust and unfair trade practices, published a framework of privacy best practices for implementing privacy and data security for companies that collect and use consumer data [25]. This framework specifies ‘unfair’ and ‘deceptive’ practices as described by Sect. 5 of the FTC Act. The Commission takes action against companies ‘that promised consumers a certain level of security (in their privacy policies, for example) and then failed to deliver’ [4]. Another example of data protection laws is Japan’s Act on the Protection of Personal Information (APPI) [26], which was fundamentally revised in both 2017 and 2022 [27]. The APPI is partially similar to relevant EU laws, especially regarding the implementation of adequate security measures, in order to ease data transfers between Japan and the EU. Overall, it has a slightly narrower scope [27].

In Europe, privacy by design is explicitly required by the GDPR, the landmark regulation governing privacy protection and data use. The scope of what is meant by ‘privacy’ in the GDPR’s ‘privacy by design’ is different from the colloquial use. The GDPR lays out a number of ‘core principles’ beyond privacy (Article 5 GDPR), in the protection of which lies its *raison d’être*. The principles with a particularly high relevance for GWAS are data protection and security, data self-determination and data fairness. The method of privacy by design (anchored in Article 25 GDPR) to protect the aforementioned principles is an obligation for systems that process personal data, which in turn is defined in

Article 4 (1) GDPR as ‘any information relating to an identified or identifiable natural person (“data subject”)’. Genomic data therefore always constitutes personal data, since it is unique to each person (and thereby identifying) even if all other identifying information (e.g. name or address) is removed [10]. In practice, pseudonymised genomic data—and subsequently the study results concerning this data subject—can generally only be matched to a person whose genomic data are both accessible and linked to them, unless re-identification through relatives’ records in online genealogy services is possible—e.g. because they entered it into a database for ancestry services. This fact lowers the identification risks associated with genomic data. But the researchers cannot simply trust that the genomic data will not be linked to a natural person either. In light of this, the rapid rise of companies and business models that sell genetic data (e.g. for forensic analyses) directly to consumers raises new questions about data protection and ethics [5, 9, 28, 29]. Privacy can never be fully ensured and the consequences can as of yet not be fully anticipated. How real these risks of leaking genetic data are is shown, for example, by last year’s successful hacking attack that exposed 6.9 million users of the ancestry service 23 and Me [30], which resulted in a class action lawsuit against 23andMe for negligence and violation of the Illinois and California law [31]. The class lawsuit is based on allegations that the company failed to take reasonable security measures to protect its customers’ sensitive data. If the class action is successful, the damage could amount to between 1 and 2 million dollars [32]. Similarly, violations of the GDPR may result in high fines or damages claims (Article 82 GDPR).

Current practices and their legal issues

In our assessment of the compatibility of current GWAS practices with privacy by design requirements, we examine a number of legal issues that need to be addressed for the various data processing steps of a typical GWAS analysis. Especially relevant to GWAS practitioners are the legal challenges arising from the core principles of the GDPR: namely data protection and security, data self-determination and data fairness, all of which must be ensured through privacy by design. Figure 1 provides a general overview of the principles. Subsequently, we address specific challenges in a GWAS context.

Firstly, the exchange of genetic data is risky from a data protection and data security perspective, as individuals are identifiable by their genetic data (genetic fingerprint). This comes with a number of challenges that are (also) relevant in a GWAS context, of which we will explain four in more detail here:

1. Technical and organisational measures

Researchers, who are usually the party controlling the data (according to the GDPR: the ‘controller’ see Article 4 (7), Article 24 GDPR), must take ‘appropriate technical and organisational measures’ (Article 25.1 GDPR) to ensure data privacy and protection and minimise the risk of data breaches (i.e. accidental or unlawful destruction, loss or unauthorised disclosure of personal data) [13]. Due to the sensitivity of genomic data, data security should be embedded as an operating principle in the organisation (akin to a ‘safety first’ culture), and technical measures such as encryption and authentication/authorisation must be robustly implemented.

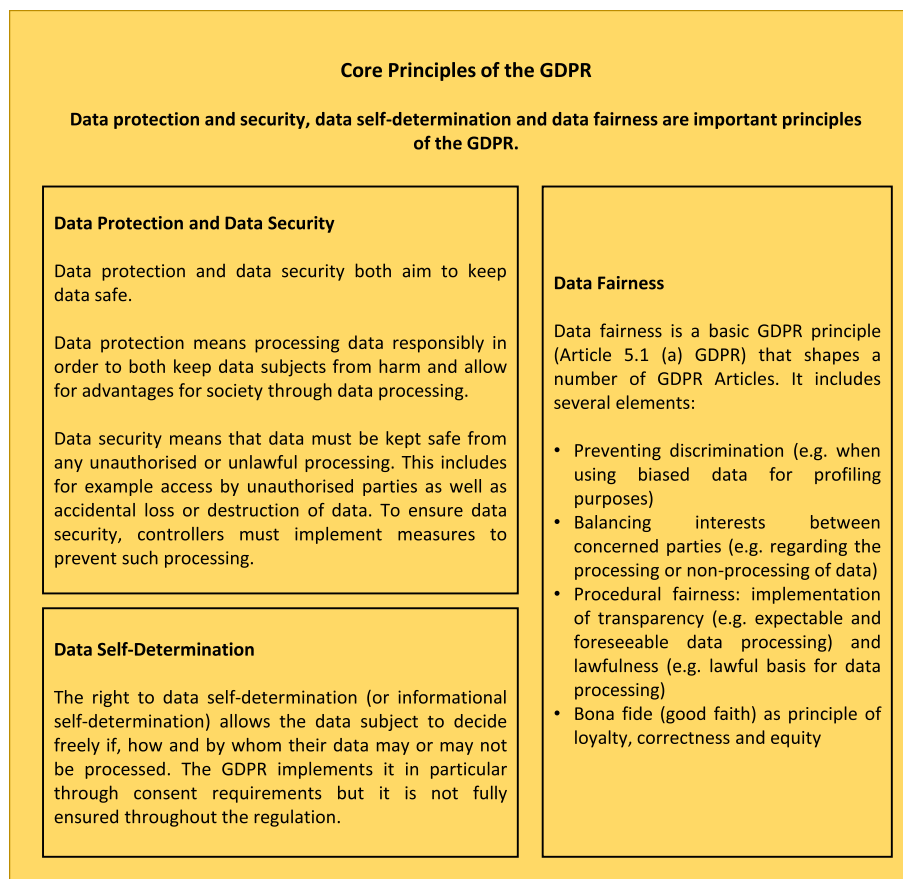


Fig. 1 Core principles of the GDPR: overview over the GDPR principles of data protection and security [33] data self-determination [34] and data fairness [35] that have to be fulfilled by GWAS researchers

Some variants of GWAS approaches already include safety measures such as homomorphic encryption (HE) in their initial set-ups [36, 37]. And with regard to authentication/authorisation, trusted research environments (TREs) are an often-used option to prevent unauthorised access to de-identified data and/or re-identification of individuals [38]. A difficulty here that leads to legal challenges is that many research institutions and data providers use their own TREs for analysis purposes, so the data are often kept separately: Even if researchers have permission to use data from two separate TREs via multi-party TREs, it is often challenging to combine the data sets [39]. The reason for this are data use agreements that have to be negotiated. Measured by the size of the data set, the sensitivity of the data and the number of people who should have access to the data, these agreements are complex, time-consuming and therefore expensive.

Additionally, the necessary security standard is kept vague by both legislation and courts and has to be determined on a case-by-case basis which makes it difficult for practitioners to establish and adopt adequate security standards.

2. Security duration

Another data protection and security challenge is that personal data must be kept secure either until its deleted or for at least the duration of the data subject's life [40], if not for that of close family members. The latter could be the case for genomic data: they differ from other personal data as they are directly linked to more than one person. No final decision on the status and rights of family members under the GDPR has been reached so far, but some scholars make strong—if controversial—cases that the need for data security does not diminish with the data subject's death as far as the data reveals information about their relatives [41, 42].

3. Cross-border transfers

Depending on the location where research is to be conducted, additional difficulties for appropriate data protection arise from cross-border legislation transfers. This is particularly relevant for GWAS that are conducted in the EU and rely on the use of genotype imputation servers located in the USA [43, 44]. Imputation is used in almost every meta- or single GWAS study to combine data from different research sites and from different array/sequencing experiments. In this step of a GWAS, the data are still identifiable (Fig. 3, Step 3), and locally performed imputation by data protection-friendly genotype imputation servers located in the EU [45] is not always feasible. Regarding GWAS conducted in the USA, cross-border transfers are necessary if the study relies on data from EU subjects.

Two adequacy decisions by the European Commission, the so-called Safe Harbor Agreement and the so-called EU-U.S. Privacy Shield, have so far failed to provide a sufficiently secure basis for data transfers to the USA and were both declared invalid by the Court of Justice of the European Union (CJEU) (2015 Schrems I judgement [46] and 2020 Schrems II judgement [47]). Since July 10, 2023, the third adequacy decision, the so-called EU-U.S. Data Protection Framework (DPF), has been in force, covering all data transfers between the EU and the USA. This new adequacy decision will allow the transfer of personal data from the EU to the USA without the need for additional safeguards such as standard contractual clauses. To apply, it requires recipients in the USA to 'join the DPF by committing to the DPF principles and self-certifying with the U.S. Department of Commerce' [48]. The majority of public sector entities in the USA, as well as banks, airlines and insurance companies, are exempt from certification and therefore do not fall under the framework [49]. Data transfers to non-DPF-certified recipients require other safeguards in accordance with Article 46 GDPR (e.g. standard contractual clauses) [48, 50]. It remains to be seen whether the new adequacy decision will once again be challenged before the CJEU. The first private action to have the data protection framework agreement annulled was dismissed by the General Court of the European Union at the beginning of October last year. To our knowledge, the relevant U.S. imputation servers are not yet DPF-certified. For this reason, GWAS researchers who want to utilise U.S. imputation servers do not benefit from the advantages, in particular the intended legal certainty, that arise from the DPF. International imputation currently remains a data processing procedure that is legally complicated and often time-consuming. In lieu of

the DPF, Article 46 GDPR mandates that appropriate safeguards must be taken and the European Commission published new standard contractual clauses in June 2021, which are mandatory for new contracts from 27 September 2021 [51]. This option requires more effort and time and lacks the benefit of legal certainty as to what constitutes appropriate safeguards that the DPF offers.

Furthermore, cross-border transfers require researchers to consider two legislations. Even though the GDPR is currently one of the strictest privacy laws in effect, it naturally does not cover every data protection and security provision under other jurisdictions.

4. Imputation methods

Imputation usually necessitates a data transfer to a third party. This leads to additional security risks. One way to guarantee such an adequate level of protection is provided by privacy-friendly genotype imputation methods. An example of such a privacy-friendly imputation method is *p-Impute*, which is based on HE [52]. *P-Impute* users can perform genotype imputation on encrypted genotype data and receive encrypted genotype outputs. A downside is that although the *p-impute* algorithm is faster due to the lack of a phasing step, it leads in its current form to lower accuracy for heterozygous SNPs [52]. Another HE-based method was presented by Kim et al. [53]. A comparison with state-of-the-art non-secure methods showed that HE-based solutions achieved comparable accuracy for common variants, but not for rare variants. An alternative to these HE-based frameworks are privacy-preserving imputation services based on trusted execution environment (TEE) technology, for example Intel SGX [54]. Due to the fact that it is hardware-based, the computational overhead is relatively small, as most of the computation is performed on the basis of the plaintext data inside the enclave, resulting in state-of-the-art imputation accuracy, which was significantly higher than HE-based solutions [54]. However, hardware-based solutions are not a homogenous concept in terms of trustworthiness [52, 55, 56] so they still often rely on users trusting the service provider to process sensitive data securely, which is not required with HE-based solutions [52]. They furthermore sacrifice some safety guarantees, which means that they do not have 'the mathematically provable safety guarantees of HE' [55]. For further details on HE [5, 57] and other PETs [29, 58, 59], see Fig. 2.

Secondly, the participants' data self-determination must be protected, especially in the form of consent. The GDPR creates several requirements for gaining consent for the processing of health and genetic data (Article 9.1 GDPR) and implementing measures to ensure the security of processing (Articles 24, 25, 32 GDPR). As a result, it is generally prohibited to process health and genetic data. The most prevalent exception to this rule is explicit consent (Article 9.2 lit. a GDPR). Consent management in GWAS though is becoming increasingly difficult due to the ever-growing number of participants in GWAS studies, with millions of participants already [64]. This becomes especially apparent in studies obtaining their data from biobanks. These may rely on very broadly worded consent forms to be effective and competitive, depending on the specific biobank collection—departmental collections, project-specific collections or hospital-wide collection [65, 66]. However, for compliance, consent must be very concrete and precise, explicitly permitting the use of genetic data and outlining

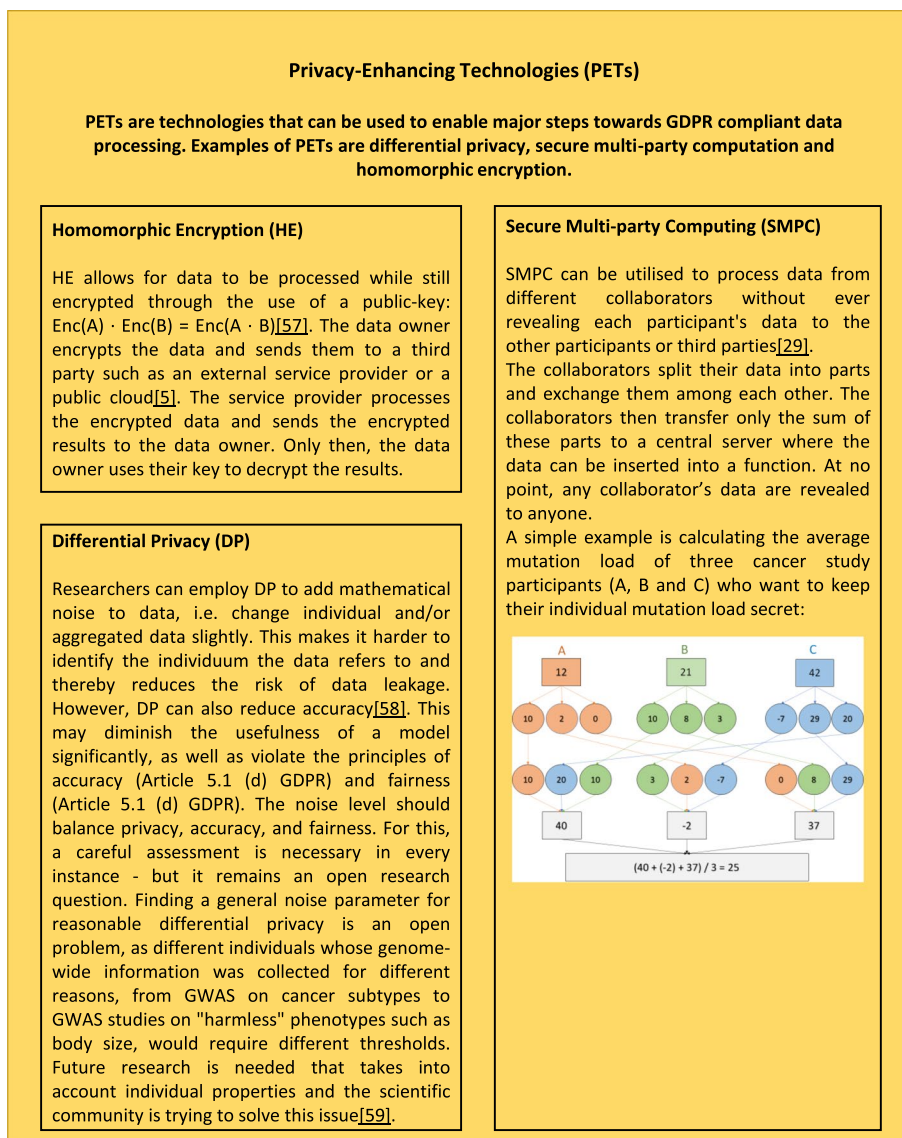


Fig. 2 Overview of the function and aim of the three standard PETs: DP [60], SMPC [61] and HE [62] are three PETs that can—depending on circumstances alone or in combination—be used to fulfil privacy by design requirements. PETs can help to protect informational self-determination by ensuring that no unauthorised parties gain access to personal data. However, other data protection requirements, such as the principle of data fairness, are largely unaffected by PETs and must be ensured separately [63]

the circumstances and possible future processing changes (as far as foreseeable in scientific terms). The concept of broad consent, that would allow for data to be processed in the context of yet unspecified projects and thereby allows for secondary use of data without a need for repeated consent, therefore generally conflicts with this need for specificity according to Article 5.1 lit. b GDPR [67]. Whilst recital 33 of the GDPR generally allows less strict requirements that would permit broad consent [67], the Article 29 Working Party (an independent advisory body to the European Commission) has signalled the prevalence of a stricter standard in 2017 (Working Party's Guidelines on consent under Regulation 2016/679) [67]. However, broad consent is

widely accepted by patients if it has been communicated as part of patient counseling [68]. Nevertheless, most patients do not wish to lose all control over their data and would prefer to make a new decision if research or processing circumstances change [66]. In any case, consent—including broad consent—needs a suitable information basis to be legitimate [69]. For this reason, patients should be informed as specifically as possible, in particular about whether ‘data is going to be shared with other research partners and across national borders,’ whether linkage to registry data is to take place or whether research results or incidental findings are to be reported back [69]. Additionally, further complications arise from the fact that genomic data also always includes information about parents and close relatives, in particular in the case of monozygotic twins due to identical DNA, leading to yet unsolved issues regarding consent. At the moment, researchers can only achieve legal certainty if federal or state laws permit the processing of the data without consent [42] or the data processing is based on another processing basis in accordance with Article 9.2 GDPR.

Data can only be processed as long as valid consent exists. A later withdrawal of consent does not, however, affect the lawfulness of processing based on consent before its withdrawal (Article 7 (3) GDPR). Analyses carried out at the time of consent can thus continue to be used lawfully despite withdrawal. However, from the time of revocation, it is unlawful for parties to request the raw data for the purpose of verification or review. Consequently, researchers must take special legal and organisational measures to protect the participants’ (data) rights (see, for example: Politou et al. [70]).

In addition to consent, the self-determination right is also protected through transparency requirements (Articles 12–15 GDPR). An issue that may arise specifically in the context of GWAS and on which EU legislators have yet to make a decision is the right of a person to know or not to know about incidental genetic findings, i.e. cases in which scientists encounter genetic variants in their studies that affect a disease other than the one being studied [71]. Generally, this decision should be left to the data subject and is often also asked for when broad consent is obtained [68]. The question of how to deal with incidental findings is typically important when dealing with rare genetic variants (i.e. genetic variants with a low frequency in the population under study but high estimated risk of disease) or mutations in genes that are known to have a major impact on the development of a disease (e.g. genetic mutations in the breast cancer genes BRCA1 and BRCA2 usually have a major influence on the development of breast cancer). Information on incidental findings is therefore also particularly sensitive and must be protected.

Thirdly, genetic discrimination, violating the principle of data fairness, can occur when apparently population-specific risk factors are identified or when they incorrectly lead to systematically biased (discriminatory) results for a particular population group (many GWAS studies listed in the GWAS Catalog of the National Human Genome Research Institute (NHGRI) [72] predominantly contains data on white populations) due to insufficient data precision for other ethnicities, for example in polygenic risk scores [73], which can, for example, lead to false prognoses [74, 75]. Furthermore, measures such as DP may skew the data leading to similar results or reinforcing bias [76]. To counteract this, the risks of DP have to be carefully considered before any amount of noise is added, and studies increasingly involve more specific populations from different parts of the

world [74] or clearly point out the limitation of the study, namely that conclusions for other populations should be considered with caution [64].

Comparison of the current genome-wide association study designs with regard to their privacy by design compatibility

In the following, we outline three different types of GWAS study designs and highlight certain special characteristics that need particular attention from a privacy by design perspective. The approaches differ in their data security and statistical power, especially in their robustness to data heterogeneity due to the heterogeneous nature inherent in biomedical data: (1) the centralised GWAS approach with all genotype data from different study populations pooled at one analysis site, (2) the meta-analysis approach, in which the genotypes from all participating studies are first analysed individually at participating centres and then only GWAS summary statistics are shared and meta-analysed and (3) the newly proposed federated analysis approach, in which genotype data stored separately at each institution is used to train local machine learning models which are then aggregated into one *federated* global model. Figure 3 provides a summary over the six typical GWAS data processing steps and inherent privacy and accuracy risks.

The responsibility to ensure privacy by design begins at the latest when researchers gain access to the genotype data (either at the quality control (step 2, see Fig. 3) or later) and continues for all processing steps where the data uniquely identify the data subjects. This is the case until the data are aggregated after single-nucleotide variant (SNV) testing and the retaining of any individual, privacy-sensitive information can be discarded (for meta-analysis and federated analysis approaches, this applies after step 4, see Fig. 3). The comparatively weaker privacy protection is a notable downside of central data processing (centralised analysis), which applies here up to step 6, see Fig. 3, if final figures or descriptive statistics are still produced using genotype data.

Since genomic data are inherently identifying, pseudonymisation alone is not sufficient to protect the rights of data subjects during these steps [11]. Therefore, researchers should additionally delete unnecessary and unusable data (e.g. local sample identifier) as soon as possible and deploy PETs such as DP, HE and SMPC to ensure secure communication and counter typical cyberattack schemes by design (see Fig. 2 for further details). It is crucial to evaluate and balance the trade-off between accuracy, computation time and data security due to the use of PETs carefully before choosing to apply—or forgo—any particular method [16, 58, 78]. If researchers utilise an already-established database (e.g. the UK Biobank [79]), the safe storage (as well as deletion after the end of the project) of the extracted data must be assured by the responsible third party, because individual study participants (e.g. from the UK Biobank) also regularly withdraw their consent.

Whilst aggregating individual data on centralised analysis servers is desirable from a research perspective, it massively increases data security risks (see Fig. 3, ‘loss of privacy’). Hence, separation of genotype data in distributed data silos is recommended from a legal point of view. Storing genomic data on large central servers also carries the risk of this data being stolen by hackers because in the event of a successful attack, a large amount of genomic data from a large number of individuals falls into the hands of the attacker all at once. There are two possible approaches suitable for conducting

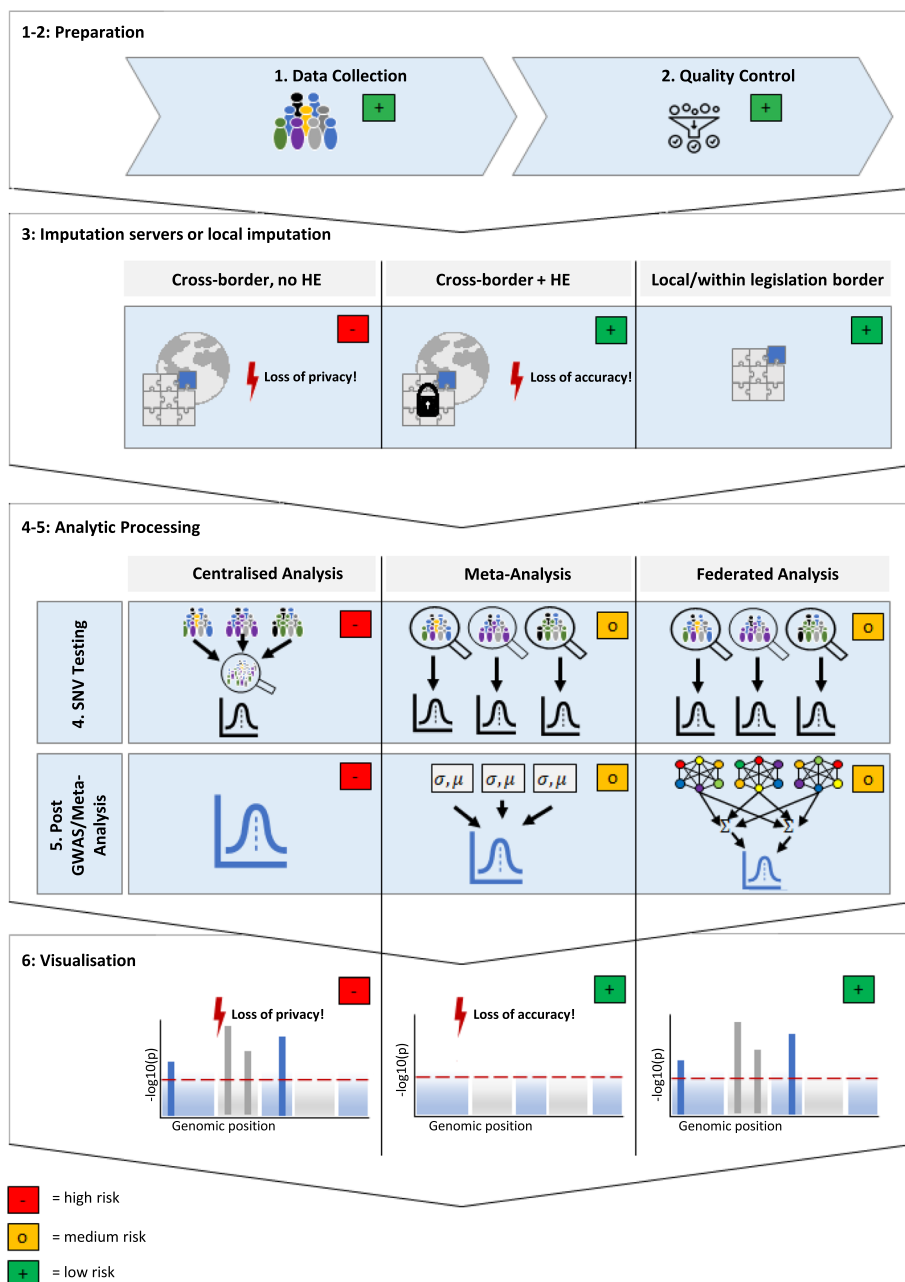


Fig. 3 Graphical summary of typical GWAS data processing steps, inherent privacy risks and risks of reduced accuracy for centralised, meta and federated analysis. Data privacy during data collection (1) and quality control (2) can be ensured through the implementation of PETs, such as HE, DP and SMPC [77] and strict access control. There are no accuracy concerns. Data privacy standards during imputation (3) differ between international imputation servers and servers imputing locally or within a legislation border, such as EagleImp-Web (EU) [45], the Haplotype Reference Consortium (UK) [43] and TopMed [44]. Whereby local imputation can lead to a loss of accuracy in the case of obsolete references and algorithms, cross-legislation border transfers lead to additional privacy risks that can be combated through the use of HE, e.g. through the tool *p-Impute*. However, this still results in lower accuracy. Privacy risks and accuracy during the last three steps (4–6) depend on the GWAS study design: a centralised analysis is accurate but leads to privacy risks. A meta-analysis has a medium to low privacy risk but may suffer from reduced accuracy. A federated approach combines a medium to low privacy risk and high accuracy

GWAS analyses on such distributed data: meta-analysis and federated GWAS. Meta-analysis aggregates the summary statistics for genomic loci for each of the distributed datasets, avoiding direct sharing of genomic data. However, this may come at the cost of accuracy for data sets with heterogeneously distributed class labels and confounding factors (e.g. uneven distribution of smokers across the data centres or unbalanced case–control ratios) [21, 80, 81] (see Fig. 3, Step 6). As mentioned at the beginning, this can lead to discriminatory outcomes which can, as well as exposing practitioners to legal risks, create real-world harm. Federated GWAS work through the emerging technology of federated learning, where statistical models are trained locally at each data centre, followed by a subsequent (or iterative) exchange of the model parameters either with a central server or with the other partners in a server-free manner, e.g. using SMPC, to produce a single joint model without exchanging the genotype data [21, 82]. Despite heterogeneously distributed phenotypes or confounding factors in different cohorts, in principle, the same results can be obtained as with centralised analysis, thus satisfying the requirements for accuracy [21] (see Fig. 3, step 6). An example for the accuracy of federated analysis was provided by Froelicher et al., whose Secure Federated Principal component analysis (SF-PCA) algorithm combines multiparty homomorphic encryption, interactive protocols and edge computing [83].

Whilst, as mentioned above, statistical scores such as the results of GWAS analyses generally do not have (directly) identifying qualities, there is a residual risk of revealing information about individual subjects; both the summary statistics in meta-analysis and the exchanged model parameters of federated GWAS may therefore in theory constitute personal data—in which case they would also require very stringent protection (although to a much lesser extent than the genomic data in centralised analysis). There is an ongoing debate in the scientific community if and to which degree one could re-identify an individual from a meta-GWAS' summary statistics if, e.g. one knows a few hundred SNPs from that individual [16]. A simulated statistical attack showed that the presence of an individual in a GWAS cohort could be determined on the basis of the aggregated allele frequencies, provided the attacker has access to some raw genomic information about the individual in question [10, 15, 16, 84, 85]. The risk grows with the increasing collection of and access to (sometimes public) personal genetic marker data [85]. Another study demonstrated that genetic data can be matched with photographs, a risk that can be addressed by using DP for images [86]. According to the study, without access to high-quality, preferably three-dimensional images, the risk is small but not negligible, especially given the ever-developing camera and artificial intelligence (AI) technologies [86].

These findings make it even more important to implement safeguards to keep genomic data secure, especially when third parties (albeit exclusively for the intermediate storage of genomic data) are given access to provide services such as genotype imputation. Effective measures to mitigate such an attack on the trained statistical model itself include employing additional PETs that either add noise to the data (DP), specifically obfuscate the underlying data (by altering carefully selected linkage disequilibrium data) [16], generalise the data (by replacing values with general but semantically consistent ones) that suppress identifying data by removing specific values or by detecting and removing outliers before model training. These measures must be

Recommendations for Researchers

Throughout a distributed GWAS analysis, appropriate PETs should be employed to reduce the risk of successful identification of data subjects. Additionally, we recommend using the following steps regarding the processes shown in Figure 3. However, it is important to note that complete data protection and fully exploiting the potential of processing sensitive data can rarely be reconciled. In this respect, a compromise must be found for each individual case that succeeds in balancing GDPR requirements and accuracy:

1. **Data collection:** If genotype data collection involves consolidation of electronic medical records (EMRs) from multiple institutions, privacy-preserving record linkage (PPRL) techniques could be used[87]. This is not yet practised in most GWAS but is likely to be of interest in the future as recent GWAS studies are increasingly based on EMRs and related biobanks of patient collectives.
2. **Data storage:** Federated data platforms are emerging as important resources to facilitate the secure exchange of data without the need to physically move the data outside its organisational or legal boundaries. For this purpose, for example, multi-party TREs have been developed[39], which also provide a safe space for data analysis[88].
3. **Quality control:** Apply quality control tools locally. In cases this is not possible, state of the art privacy practices should be observed and additional safety measures implemented (see, for example, secure private clouds[89] such as computerome[90]). So-called cookbooks can be helpful in this context, and a generation of specially coded and aggregated statistics (so-called partial derivations) for secure predefined association tests (see point 5) can already take place[91], so that no person-level information can be obtained from the genotype matrices after coding.
4. **Imputation:** Apply local tools if possible or federated solutions (e.g. p-Impute)[52]. In many cases, this is infeasible because some imputation reference panels are only accessible through the use of proprietary servers. Especially in cases where data that falls under the scope of the GDPR is processed and the server is outside the EU (e.g. TOPMed in the US, or HRC Sanger Imputation Service in the UK) this leads to privacy concerns. Researchers must balance the privacy and effectiveness of each option. One option for data processing that is governed by the GDPR would be to use GDPR-compliant phasing and imputation web services, which are now also being set up in the EU[45].
5. **SNV association testing and follow-up analysis:** If the desired statistical tests are already implemented as a “federated” version, use federated GWAS tools (e.g. sPLINK or iPRIVACY)[21].
6. **Visualisation:** PETs are not required if only GWAS summary statistics are used, so that the use of standard tools on the statistical model output (e.g. for producing Manhattan plots) is possible.

Fig. 4 Overview of recommendations for researchers in relation to data collection, data storage, quality control, genotype imputation, SNV association testing and follow-up analysis and visualisation in distributed GWAS analysis

carefully chosen and administered to balance the accuracy-privacy trade-off in a way that is suitable for the sample populations’ specific features, such as demographics or outliers [59, 78] (see Fig. 4 for further details). Researchers must carry out this complex balancing process (e.g. have to choose the appropriate level of noise); in doing so, they are not bound by any specific legal requirements. Rather, they must protect the data in a way that corresponds especially to the state of the art, the probability of occurrence and the severity of the risks associated with the processing for the rights and freedoms of natural persons (Article 25.1 GDPR). At the same time, they must take into account technical developments that lead to both more secure measures and new risks, if they can be considered at all based on the characteristics of the study, to achieve the intended effect. In some cases, it might also be required to combine PETs to heighten their effectiveness and compensate for weaknesses. This especially pertains to DP as the amount of noise necessary to achieve adequate data privacy in the context of a GWAS is not reasonable with respect to accuracy concerns. This can

be circumvented by combining DP with HE [5] or SMPC [36]. HE and SMPC profit in return from reduced communication and computation overhead [36]. Figure 4 shows which concrete actions can be taken when projects fall within the scope of the GDPR in the following six areas: data collection [87], data storage [39, 88], quality control [89–91], genotype imputation [45, 52], SNV association tests and follow-up analysis [21] as well as visualisation in distributed GWAS analysis.

Conclusion

Fully ensuring privacy by design in GWAS comes with a number of challenges, but researchers who are conscious of these challenges and wish to tackle them head-on have access to a growing array of methods and tools. In particular, federated GWAS has the potential to overcome the most persistent privacy challenges for GWAS on distributed datasets and multi-centred GWAS meta-analysis whilst avoiding unacceptable accuracy trade-offs. However, it is hampered by legal uncertainties and a number of (yet) unresolved legal questions. The four most important ones which we discussed in our contribution are cross-border data transfers for genotype imputation purposes, the rights of family members regarding genomic data shared for study purposes, the legality of consent and a lack of diversity in studies of populations leading to genetic discrimination (or sometimes opens up discrimination, when genetic variation is studied only for certain population groups) [75]. Interestingly, contrary to previous assumptions, published GWAS summary statistics as stored in public databases such as the NHGRI GWAS Catalogue can reveal an individual's participation in trait-specific GWAS. In the case of GWAS for diseases, this may lead to unwanted and unauthorised disclosure of information. For these and an array of further technical questions, guidelines and ultimately a robust legal framework are needed to provide legal certainty and improve compliance for the handling of genomic data in research under the GDPR regulatory regime. Data protection and privacy risks cannot be completely eliminated, but can be largely combated by implementing PETs and processing data locally where appropriate—these are fundamental aspects (along with self-determination and discrimination prevention) of privacy by design.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-024-03296-6>.

Supplementary Material 1.

Review history

The review history is available as Additional file 1.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

Conceived of the review, AB and LS. Writing, AB, LS, SW, GB, JB and DE. Figures, AB and LS. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. Our work was funded by the German Federal Ministry of Education and Research (BMBF) (*grants 16DTM100A and 16DTM100C*). This project has also received funding from the European Union's Horizon2020 research and innovation programme under grant agreement no. 826078. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. The project received infrastructure support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Cluster of Excellence 2167 'Precision Medicine in Chronic Inflammation (PMI)' (EXC 2167–390884018) and the DFG research unit 'miTarget' (project number 426660215; INF (EL 831/5–2)). We further acknowledge financial support from the Open Access Publication Fund of Universität Hamburg.

Declarations

Ethics approval and consent to participate

Ethical approval is not applicable for this article.

Competing interests

The authors declare no competing interests.

Received: 16 May 2023 Accepted: 3 June 2024

Published: 13 June 2024

References

1. General Data Protection Legislation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC Apr 27, 2016. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.
2. California Legislative Information. California Consumer Privacy Act of 2018. Available from: https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=2017201805B1121.
3. Shabani M, Borry P. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. *Eur J Hum Genet.* 2018;26:149–56.
4. Pardau SL, Edwards B. The FTC, the unfairness doctrine, and privacy by design: new legal frontiers in cybersecurity. *J Business Technol Law.* 2017;12:227–76.
5. Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. *Nat Rev Genet.* 2022;23:429–45.
6. Bednar K, Spiekermann S, Langheinrich M. Engineering privacy by design: are engineers ready to live up to the challenge?. *arXiv [cs.CY].* 2020. Available from: <http://arxiv.org/abs/2006.04579>.
7. Berger B, Cho H. Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biol.* 2019;20:128.
8. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nat Rev Genet.* 2014;15:409–21.
9. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. *Nat Genet.* 2020;52:646–54.
10. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR: assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. *EMBO Rep.* 2019;20:e48316. <https://doi.org/10.15252/embr.201948316>.
11. Colin Mitchell, Johan Ordish, Emma Johnson, Tanya Brigden and Alison Hall. The GDPR and genomic data. PHG Foundation; 2020 May. Available from: <https://www.phgfoundation.org/report/the-gdpr-and-genomic-data>.
12. Quinn P, Quinn L. Big genetic data and its big data protection challenges. *Comput Law Secur Rev.* 2018;34:1000–18.
13. Brauneck A, Schmalhorst L, Kazemi Majdabadi MM, Bakhtiari M, Völker U, Baumbach J, et al. Federated machine learning, privacy-enhancing technologies, and data protection laws in medical research: scoping review. *J Med Internet Res.* 2023;25:e41588.
14. Wang X, Dervishi L, Li W, Ayday E, Jiang X, Vaidya J. Privacy-preserving federated genome-wide association studies via dynamic sampling. *Bioinformatics.* 2023;39:btad639. <https://doi.org/10.1093/bioinformatics/btad639>.
15. Homer N, Szelling S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *Plos Genet.* 2008;4:e1000167.
16. Wang R, Li YF, Wang X, Tang H, Zhou X. Learning your identity and disease from research papers: information leaks in genome wide association study. *Proceedings of the 16th ACM conference on Computer and communications security.* New York, NY, USA: Association for Computing Machinery; 2009. p. 534–44.
17. Humbert M, Ayday E, Hubaux J-P, Telenti A, Telenti A. Addressing the concerns of the lacks family: quantification of kin genomic privacy. *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security.* New York, NY, USA: Association for Computing Machinery; 2013. p. 1141–52.
18. Mizas C, Sirakoulis GC, Mardiris V, Karafyllidis I, Glykos N, Sandaltzopoulos R. Reconstruction of DNA sequences using genetic algorithms and cellular automata: towards mutation prediction? *Biosystems.* 2008;92:61–8.
19. Bossé Y, Amos CI. A decade of GWAS results in lung cancer. *Cancer Epidemiol Biomarkers Prev.* 2018;27:363–79.
20. Constable SD, Tang Y, Wang S, Jiang X, Chapin S. Privacy-preserving GWAS analysis on federated genomic datasets. *BMC Med Inform Decis Mak.* 2015;15(Suppl 5):S2.

21. Nasirigerdeh R, Torkzadehmahani R, Matschinske J, Frisch T, List M, Späth J, et al. sPLINK: a federated, privacy-preserving tool as a robust alternative to meta-analysis in genome-wide association studies. *bioRxiv*. 2022. p. 2020.06.05.136382. Available from: <https://www.biorxiv.org/content/10.1101/2020.06.05.136382v2>. Cited 2022 Aug 2.
22. Psychiatric Genomics Consortium. Available from: <https://pgc.unc.edu/about-us/>. Cited 2023 Feb 15.
23. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511:421–7.
24. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PIW, Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. 2007;316:1331–6.
25. Federal Trade Commission. Protecting Consumer Privacy in an Era of Rapid Change. Federal Trade Commission; 2012 Mar. Available from: <https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf>.
26. Act on the Protection of Personal Information - English - Japanese Law Translation. Available from: <https://www.japaneselawtranslation.go.jp/en/laws/view/2781/en>. Cited 2023 Feb 15.
27. González G, Van Brakel R, De Hert P. Research handbook on privacy and data protection law: values, norms and global politics. Cheltenham: Edward Elgar Publishing; 2022.
28. Regalado A. More than 26 million people have taken an at-home ancestry test. *MIT Technology Review*. 2019. Available from: <https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/>. Cited 2024 Jan 30.
29. Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux J-P, et al. Privacy in the genomic era. *ACM Comput Surv*. 2015;48:1. <https://doi.org/10.1145/2767007>.
30. Carballo R. Data Breach at 23andMe Affects 6.9 Million Profiles, Company Says. *The New York Times*. 2023. Available from: <https://www.nytimes.com/2023/12/04/us/23andme-hack-data.html>. Cited 2024 Jan 31.
31. Bucher A. 23andMe hit with another class action lawsuit over data breach. *Top Class Actions*. 2023. Available from: <https://topclassactions.com/lawsuit-settlements/privacy/data-breach/23andme-hit-with-another-class-action-lawsuit-over-data-breach/>. Cited 2024 Jan 31.
32. Jon Styf AJ. 23andMe reportedly blames data breach on victims. *Top Class Actions*. 2024. Available from: <https://topclassactions.com/lawsuit-settlements/privacy/data-breach/23andme-confirms-oct-breach-compromised-data-from-6-9m-users/>. Cited 2024 Jan 31.
33. Pinheiro PP, Battaglini HB. Artificial intelligence and data protection: a comparative analysis of AI regulation through the lens of data protection in the EU and Brazil. *GRUR Int*. 2022;71:924–32.
34. Thouvenin F. Informational self-determination: a convincing rationale for data protection law? *J Intell Prop Info Tech & Elec Com L*. 2021;12:246–56.
35. Malgieri G. The concept of fairness in the GDPR: a linguistic and contextual interpretation. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery; 2020. p. 154–66.
36. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nat Commun*. 2021;12:5910.
37. Blatt M, Gusev A, Polyakov Y, Goldwasser S. Secure large-scale genome-wide association studies using homomorphic encryption. *Proc Natl Acad Sci U S A*. 2020;117:11608–13.
38. Sudlow C. Trusted Research Environments. *HDR UK*. 2021. Available from: <https://www.hdr.ac.uk/access-to-health-data/trusted-research-environments/>. Cited 2023 Feb 13.
39. Waind E. Multi-party trusted research environment federation: Establishing infrastructure for secure analysis across different clinical-genomic datasets. *DARE UK*. 2022. Available from: <https://dareuk.org.uk/multi-party-trusted-research-environment-federation-clinical-genomic-datasets/>. Cited 2023 Feb 13.
40. Buchmann J, Geihs M, Hamacher K, Katzenbeisser S, Stammler S. Long-term integrity protection of genomic data. *EURASIP J Inf Secur*. 2019;2019:1–14.
41. Kuru T. Genetic data: the Achilles' heel of the GDPR? *Eur Data Prot Law Rev*. 2021;7:45–58.
42. Kuru T, de Beriaín IM. Your genetic data is my genetic data: unveiling another enforcement issue of the GDPR. *Comp Law Sec Rev*. 2022;47:105752.
43. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48:1279–83.
44. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*. 2021;590:290–9.
45. Wienbrandt L, Prieß C, Kässens JC, Franke A, Uhing F, Ellinghaus D. EagleImp-Web: a fast and secure genotype phasing and imputation web service using field-programmable gate arrays. *bioRxiv*. 2022. p. 2022.02.24.481790. Available from: <https://www.biorxiv.org/content/10.1101/2022.02.24.481790v1>. Cited 2022 Oct 6.
46. Judgment of the Court (Grand Chamber) of 6 October 2015 (Schrems I). Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62014CJ0362>. Cited 2023 Nov 14.
47. Judgment of the Court (Grand Chamber) of 16 July 2020 (Schrems II). Available from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62018CJ0311>. Cited 2022 Oct 6.
48. Marko R, Sekanina J. The new transatlantic data privacy framework. *Transatlantic Law Journal*. 2023;2:63–5.
49. Miño V. What does the Data Privacy Framework Self-Certification mean for your company?. *datenschutz notizen | News-Blog der DSN GROUP*. 2023. Available from: <https://www.datenschutz-notizen.de/what-does-the-data-privacy-framework-self-certification-mean-for-your-company-0545511/>. Cited 2024 Jan 18.
50. Phillips M. International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR). *Hum Genet*. 2018;137:575–82.
51. New Standard Contractual Clauses - Questions and Answers overview. European Commission. Available from: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/new-standard-contractual-clauses-questions-and-answers-overview_en. Cited 2024 Feb 6.

52. Gürsoy G, Chielle E, Brannon CM, Maniatakos M, Gerstein M. Privacy-preserving genotype imputation with fully homomorphic encryption. *Cell Syst.* 2022;13:173–82.e3.
53. Kim M, Harmanci AO, Bossuat J-P, Carpov S, Cheon JH, Chillotti I, et al. Ultrafast homomorphic encryption models enable secure outsourcing of genotype imputation. *Cell Syst.* 2021;12:1108–20.e4.
54. Dokmai N, Kockan C, Zhu K, Wang X, Sahinalp SC, Cho H. Privacy-preserving genotype imputation in a trusted execution environment. *Cell Syst.* 2021;12:983–93.e7.
55. Sherman MA. Paving the path toward genomic privacy with secure imputation. *Cell Syst.* 2021;12:950–2.
56. Sabt M, Achemlal M, Bouabdallah A. Trusted execution environment: what it is, and what it is not. 2015 IEEE Trustcom/BigDataSE/ISPA. New York City: IEEE; 2015. p. 57–64.
57. Heinz C, Wall N, Wansch AH, Grimm C. Privacy, GDPR, and homomorphic encryption. In: Zivkovic C, Guan Y, Grimm C, editors. *IoT Platforms, Use Cases, Privacy, and Business Models: With Hands-on Examples Based on the VICINITY Platform*. Cham: Springer International Publishing; 2021. p. 165–84.
58. Johnson A, Shmatikov V. Privacy-preserving data exploration in genome-wide association studies. *KDD.* 2013;2013:1079–87.
59. Uhlerop C, Slavković A, Fienberg SE. Privacy-preserving data sharing for genome-wide association studies. *J Priv Confid.* 2013;5:137–66.
60. Ficek J, Wang W, Chen H, Dagne G, Daley E. Differential privacy in health research: a scoping review. *J Am Med Inform Assoc.* 2021;28:2269–76.
61. Mugunthan V, Byrd D, Balch TH, Morgan JP. SMPAL: Secure Multi-Party Computation for Federated Learning. 2019; Available from: <https://www.jpmorgan.com/content/dam/jpm/cib/complex/content/technology/ai-research-publications/pdf-9.pdf>. Cited 2022 Mar 9.
62. Truong N, Sun K, Wang S, Guitton F, Guo Y. Privacy preservation in federated learning: an insightful survey from the GDPR perspective. *Computer Security.* 2021;110. Available from: <https://www.sciencedirect.com/science/article/pii/S0167404821002261>.
63. Information Commissioner's Office. Privacy-enhancing technologies (PETs). 2023. Available from: <https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies-1-0.pdf>.
64. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. *Nature.* 2022;610:704–12.
65. Metzler I, Ferent L-M, Felt U. On samples, data, and their mobility in biobanking: How imagined travels help to relate samples and data. *Big Data Soc.* 2023;10:20539517231158636.
66. Goisauf M, Martin G, Bentzen HB, Budin-Ljøsne I, Ursin L, Durnová A, et al. Data in question: a survey of European biobank professionals on ethical, legal and societal challenges of biobank research. *Plos One.* 2019;14:e0221496.
67. Hallinan D. Broad consent under the GDPR: an optimistic perspective on a bright future. *Life Sci Soc Pol.* 2020;16:1–18.
68. Richter G, Krawczak M, Lieb W, Wolff L, Schreiber S, Buyx A. Broad consent for health care-embedded biobanking: understanding and reasons to donate in a large patient sample. *Genet Med.* 2018;20:76–82.
69. Hansson MG. Striking a balance between personalised genetics and privacy protection from the perspective of GDPR. In: Slokenberga S, Tzortzatos O, Reichel J, editors. *GDPR and Biobanking: Individual Rights, Public Interest and Research Regulation across Europe*. Cham: Springer International Publishing; 2021. p. 31–42.
70. Politou E, Alepis E, Patsakis C. Forgetting personal data and revoking consent under the GDPR: challenges and proposed solutions. *J Cyber Secur.* 2018;4. Available from: <https://academic.oup.com/cybersecurity/article-pdf/4/1/tyy001/27126900/tyy001.pdf>. Cited 2022 Aug 10.
71. de Wert G, Dondorp W, Clarke A, Dequeker EMC, Cordier C, Deans Z, et al. Opportunistic genomic screening. Recommendations of the European society of human genetics. *Eur J Hum Genet.* 2021;29:365–77.
72. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* 2023;51:D977–85.
73. King A, Wu L, Deng H-W, Shen H, Wu C. Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease. *BMC Med.* 2022;20:385.
74. Haga SB. Impact of limited population diversity of genome-wide association studies. *Genet Med.* 2010;12:81–4.
75. Wauters A, Van Hoyweghen I. Global trends on fears and concerns of genetic discrimination: a systematic literature review. *J Hum Genet.* 2016;61:275–82.
76. Renieris E. Why PETs (privacy-enhancing technologies) may not always be our friends. Available from: <https://www.adalovelaceinstitute.org/blog/privacy-enhancing-technologies-not-always-our-friends/>. Cited 2024 Jan 18.
77. Jordan S, Fontaine C, Hendricks-Sturrrup R. Selecting privacy-enhancing technologies for managing health data use. *Front Public Health.* 2022;10:814163.
78. Malin B, Loukides G, Benitez K, Clayton EW. Identifiability in biobanks: models, measures, and mitigation strategies. *Hum Genet.* 2011;130:383–92.
79. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562:203–9.
80. Zolotareva O, Nasirigerdeh R, Matschinske J, Torzadehmahani R, Bakhtiari M, Frisch T, et al. Flimma: a federated and privacy-aware tool for differential gene expression analysis. *Genome Biol.* 2021;22:338.
81. Yadav P, Ellinghaus D, Rémy G, Freitag-Wolf S, Cesaro A, Degenhardt F, et al. Genetic factors interact with tobacco smoke to modify risk for inflammatory bowel disease in humans and mice. *Gastroenterology.* 2017;153:550–65.
82. Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. *Nat Biotechnol.* 2018;36:547–51.
83. David Froelicher, Hyunghoon Cho, Manaswitha Edupalli, Joao Sa Sousa, Jean-Philippe Bossuat, Apostolos Pyrgelis, Juan R. Troncoso-Pastoriza, Bonnie Berger and Jean-Pierre Hubaux. Scalable and privacy-preserving federated

- principal component analysis. IEEE Symposium on Security and Privacy. 2023; Available from: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10179350>.
84. von Thenen N, Ayday E, Cicek AE. Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics*. 2019;35:365–71.
 85. Cai R, Hao Z, Winslett M, Xiao X, Yang Y, Zhang Z, et al. Deterministic identification of specific individuals from GWAS results. *Bioinformatics*. 2015;31:1701–7.
 86. Venkatesaramani R, Malin BA, Vorobeychik Y. Re-identification of individuals in genomic datasets using public face images. *Sci Adv*. 2021;7:eabg3296.
 87. Heidt CM, Hund H, Fegeler C. A federated record linkage algorithm for secure medical data sharing. *Stud Health Technol Inform*. 2021;278:142–9.
 88. Alvarelos M, Sheppard HE, Knarston I, Davison C, Raine N, Seeger T, et al. Democratizing clinical-genomic data: how federated platforms can promote benefits sharing in genomics. *Front Genet*. 2022;13:1045450.
 89. Olowu M, Yinka-Banjo C, Misra S, Florez H. A secured private-cloud computing system. *Applied Informatics*. Madrid: Springer International Publishing; 2019. p. 373–84.
 90. Technical University of Denmark. Computerome. Available from: <https://www.computerome.dk/solutions/secure-private-cloud>. Cited 2023 Feb 27.
 91. Cookbook for eQTLGen phase II analyses - eQTLGen Phase II. Available from: <https://eqtlgen.github.io/eqtlgen-website/eQTLGen-p2-cookbook.html>. Cited 2023 Mar 16.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.