Consent Modules for Data Sharing via the German Human Genome-Phenome Archive (GHGA)

Version: 1.0, Date: 13 July 2022

Authors – The GHGA Consent Task Force:

Andreas Bruns[1*], Anna Benet-Pages[3,4], Jan Eufinger[5], Holm Graessner[6,7], Oliver Kohlbacher[8,9,10], Fruzsina Molnár-Gábor[11], Simon Parker[5], Christoph Schickhardt[2], Oliver Stegle[12,13], Eva Winkler[1]**

* Corresponding author: andreas.bruns@med.uni-heidelberg.de
** Corresponding author: eva.winkler@med.uni-heidelberg.de

[1] Section Translational Medical Ethics, National Center for Tumor Diseases (NCT), University Hospital Heidelberg (UKHD), Heidelberg
[2] Section Translational Medical Ethics, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg
[3] Bioinformatics Department, Medical Genetics Center Munich (MGZ), Munich
[4] Institute of Neurogenomics, Helmholtz Munich (HMGU), Munich
[5] German Human Genome-Phenome Archive (GHGA, W620), German Cancer Research Center (DKFZ), Heidelberg
[6] Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen
[7] Centre for Rare Diseases (ZSE) Tübingen, University Hospital Tübingen, Tübingen
[8] Applied Bioinformatics, Department of Computer Science, University of Tübingen, Tübingen
[9] Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen
[10] Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen
[11] Faculty of Law, BioQuant Centre, Heidelberg University, Heidelberg
[12] Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg
[13] European Molecular Biology Laboratory (EMBL), Heidelberg

Special thanks to:

Eric Apondo (Section Translational Medical Ethics, National Center for Tumor Diseases, University Hospital Heidelberg, Heidelberg), **Lisa Kaldowski** (Faculty of Law, BioQuant Centre, Heidelberg University, Heidelberg), **Martin Lablans** (Federated Information Systems, University Medical Center Mannheim and German Cancer Research Center, Heidelberg), **Susanne Motameny** (Cologne Center for Genomics, University of Cologne, Cologne), **Ameli Schenk** (Faculty of Law, BioQuant Centre, Heidelberg University, Heidelberg)

Table of Contents

List of Abbreviations

GHGA	German Human Genome-Phenome Archive
EGA	(Federated) European Genome-Phenome Archive
DFG	German Research Foundation
NFDI	National Research Data Infrastructure
MII	Medical Informatics Initiative
GA4GH	Global Alliance for Genomics and Health
B1MG	Beyond One Million Genomes
DKFZ	German Cancer Research Center
GDPR	General Data Protection Regulation
FAIR	Findable, Accessible, Interoperable, Reusable
DAC	Data Access Committee

Introduction

Introduction to GHGA

The German Human Genome-Phenome Archive (GHGA)[1] is a research consortium which is currently establishing a federated data infrastructure (the "GHGA Data Infrastructure") allowing the secure storage of and controlled access to omics and related health data consented for scientific research use. Moreover, GHGA strives to provide a robust ethical and legal framework balancing FAIR[2] data use and data protection concerns.

The GHGA Data Infrastructure is operated by the institutions representing the GHGA Data Hubs, with the German Cancer Research Center (DKFZ) ("GHGA Central") operating the central data hub. The project is funded by the German Research Foundation (DFG)[3] within the framework of the National Research Data Infrastructure (NFDI)[4]. GHGA also serves as a national node for the federated European Genome-Phenome Archive (EGA)[5] ensuring coordination with European and international data sharing initiatives and infrastructures.

Data Sharing via the GHGA Data Infrastructure

GHGA aims to provide a secure infrastructure for external, independent researchers and research institutions ("Data Submitters") to share their omics and related health data assets with other researchers ("Data Requesters"), nationally and internationally. Data sharing via the GHGA Data Infrastructure proceeds in four steps:

1. **Data collection:** Data Submitters obtain data from patients and/or research participants ("Data Subjects"), e.g., in the context of clinical trials or studies. The Data Submitter acts as a data controller within the meaning of the General Data Protection Regulation (GDPR)[6]. The term "Data Submitter" is thus used equivalent to "data controller" in this paper. Note that Data Submitters can also gain data controllership through various other, e.g., contractual means, without direct contact with Data Subjects.

[1] https://www.ghga.de.
[2] The FAIR Data Principles were proposed by Wilkonson et al. (The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*, 3, 2016) to help enhance the reusability of data from various sources. Along the FAIR principles, data should be made findable, accessible, interoperable, and reusable. See the full article here: https://www.nature.com/articles/sdata201618.
[3] https://www.dfg.de/en/.
[4] https://www.nfdi.de/?lang=en.
[5] https://ega-archive.org.
[6] https://gdpr-info.eu.

2. **Data submission:** Data Submitters may choose to upload datasets into the GHGA Data Infrastructure to make them available for secondary research use. GHGA Central, represented by DKFZ, serves as the data processor within the meaning of the GDPR, sharing data on behalf of and as instructed by the Data Submitter. Within GHGA, a number of other academic institutions operate Data Hubs[7] and act as sub-processors on behalf of GHGA Central.

3. **Data processing:** GHGA processes incoming data in order to ensure the quality, discoverability, accessibility, and usability of datasets within the GHGA Data Infrastructure. This happens on the basis of a contractual agreement between the Data Submitter and GHGA Central (data processing agreement).

4. **Data access:** Potential data users may browse metadata[8] on the datasets stored within the GHGA Data Infrastructure and request access to datasets of interest. Each request by a Data Requester is reviewed by the Data Submitter (or those authorised by the Data Submitter to review the request on their behalf) who then needs to establish a contractual relationship with the Data Requester (data access agreement). When notified by the Data Submitter, GHGA will transfer the data to the appropriate Data Requester, using technical and procedural control measures to ensure that the data remains secure during transfer.

Note that data sharing with the Data Requester occurs only with the fourth step. In the current model, data controllership according to the GDPR remains solely with the Data Submitter.

About This Whitepaper

Informed consent is required to share omics and related health data for scientific research use. This whitepaper contains guidance for Data Submitters on how to update their consent forms[9] such that data may be shared via the GHGA Data Infrastructure. This could also benefit comparable secure scientific research data infrastructures.

The paper contains four consent modules which together constitute the Modular Consent Toolkit:

The Data Sharing Module (Tables 1 and 2)

[7] Currently planned GHGA Data Hubs are located at the Universities of Tübingen, Munich (TUM), Cologne, Dresden, and Kiel.

[8] Metadata means data describing other data, in this case information describing the datasets within the GHGA Data Infrastructure. Metadata allows others to understand the nature of the data and may include, but is not limited to, information about the technologies used to collect the data, documentation of the Data Subject's disease status, information about the kind of consent under which data was collected, and possible restrictions on access and use.

[9] By "consent form" we mean a document that has two parts: (1) an information sheet disclosing relevant information to the Data Subject, followed by (2) a list of consent options that allows them to indicate their preferences regarding the use of their data.

The De-Identification Module (Tables 3 and 4)

The Controlled Access Module (Tables 5 and 6)

The Consent Options Module (Tables 7 and 8)

The Modular Consent Toolkit was developed as a service for Data Submitters who intend to share data for secondary research use. This may be data collected in research settings. However, genome and omics data generated for routine clinical care is another valuable resource for biomedical research. The Medical Informatics Initiative (MII)[10] aims to make clinical data collected at German university hospitals available for research and has developed a consent form template[11] to this end. GHGA's efforts should be understood as complementary to MII's efforts, and convergence and integration of the two approaches would be desirable (see Outlook).

This document and the Modular Consent Toolkit may be updated according to new knowledge or demands.[12] Our consent modules are based on a review of current best practice examples, including consent form templates such as the one produced by the MII, and the ongoing works on consent clauses and consent recommendations within the Global Alliance for Genomics and Health (GA4GH)[13] and the Beyond One Million Genomes (B1MG)[14] project.

Legal Disclaimer

As data controllers, Data Submitters are solely responsible for identifying and providing a lawful legal basis for their data processing. The Modular Consent Toolkit has been designed by GHGA members to provide guidance to Data Submitters on how to update their existing consent forms to be/remain GDPR-compliant when sharing data for scientific research purposes via the GHGA Data Infrastructure and other secure scientific research data infrastructures offering an equivalent level of data protection. It serves as an orientation for researchers and research institutions. It does not constitute formal legal advice from GHGA members, GHGA Central, or the GHGA Data Hubs.

In principle, the individual components of the Modular Consent Toolkit support Data Submitters to meet the requirements for consent pursuant to Art. 6 (1) (a), Art. 7, and Art. 9 (2) (a) of the GDPR. However, its structure is modular in that it contains general building blocks expanding on already existing consent forms of the Data Submitter and must be adapted for the specific processing context and supplemented if necessary to ensure lawful processing.

[10] https://www.medizininformatik-initiative.de/en/start.
[11] https://www.medizininformatik-initiative.de/de/mustertext-zur-patienteneinwilligung.
[12] Please check the GHGA website (www.ghga.de) for any updates or subscribe to our newsletter at: https://www.ghga.de/news-events/newsletter.
[13] https://www.ga4gh.org.
[14] https://b1mg-project.eu.

Data Submitters should always check with their data protection officer as well as their legal department whether their consent form provides a lawful legal basis for their data sharing purposes. Data Submitters are obliged to clarify compliance with the provision of purpose limitation and the clear statement of purposes for which data will be shared, to keep contact with Data Subjects and to recontact them, if necessary, in the course of application of this module, to secure the legal basis for the planned processing altogether, and to select a processor that fulfils the technical standards prescribed by the applicable data protection law.

Please note that the information to be provided in the context of the data protection consent does not automatically fulfil the information requirements pursuant to Art. 13 and Art. 14 of the GDPR.

Modular Consent Toolkit

The Data Sharing Module

Data Submitters who wish to share datasets via the GHGA Data Infrastructure should integrate the following Data Sharing Module, available both in English (Table 1) and in German (Table 2), into their consent form. If applicable, we recommend inserting a separate section on downstream data sharing into the consent form in which it is explained how data will be made available to the broader research community for research purposes beyond the primary purpose.

We would also like to deposit your pseudonymised data in the German Human Genome-Phenome Archive (www.ghga.de) or other scientific data archives offering equivalent data protection standards. This will enable us to share your data with other researchers, helping the advance of biomedical sciences, while making sure that your data is being stored securely through the application of appropriate technical and organisational measures and in compliance with European and other applicable data protection law. The archive may further process your data to ensure its scientific value and usability. Any request to access your pseudonymised data is reviewed by us beforehand and granted by the archive only after our approval. Your data may also be shared with researchers outside of the EU but only if equivalent data protection measures are in place.

Table 1: Data Sharing Module (English)

Wir würden Ihre pseudonymisierten Daten außerdem gern im Deutschen Humangenom-Phenomarchiv (www.ghga.de) oder in anderen wissenschaftlichen Datenarchiven mit gleichwertigen Datenschutzstandards speichern. So können wir Ihre Daten mit anderen Forschenden teilen, den Fortschritt der biomedizinischen Wissenschaften unterstützen und gleichzeitig sicherstellen, dass Ihre Daten durch den Einsatz geeigneter technischer

und organisatorischer Maßnahmen sicher und gemäß Europäischem und anderem anwendbarem Datenschutzrecht aufbewahrt werden. Das Archiv kann Ihre Daten weiterverarbeiten, um deren wissenschaftlichen Wert und Nutzbarkeit sicherzustellen. Jede Anfrage, um Ihre pseudonymisierten Daten zu nutzen, wird zuvor von uns geprüft und der Zugriff durch das Archiv nur nach unserer vorherigen Genehmigung erteilt. Ihre Daten können auch mit Forschenden außerhalb der EU geteilt werden, aber nur, solange dort gleichwertige Datenschutzstandards gelten.

Table 2: Data Sharing Module (German)

The Data Sharing Module describes a way of sharing data securely for scientific research that is not limited to GHGA. As such, the module mentions GHGA but would also be applicable for any other data infrastructure following the same model of secure, GDPR-compliant data sharing. Within the text the reference to equivalence means with regard to the risk posed to the Data Subject by engaging a data archive as a processor. There should not be an increase in either the likelihood or the severity of the risk posed to the rights and freedoms of the Data Subject. Especially where the consent mentions other long-term storage infrastructures, Data Submitters may wish to add that GHGA is a specialised genome archive that enables cross-border data sharing. It remains the responsibility of the Data Submitter to choose data archives that comply with the applicable data protection law.

Please note that as the module is to be integrated into an existing consent form, the terms “we”, “us”, and “our” refer to the Data Submitter (or another institution using the consent form). GHGA Central only shares the data with other researchers on behalf of the Data Submitter. If the Data Submitter reviews data access requests with the help of a Data Access Committee (DAC), they may wish to say: “Any request to access your pseudonymised data is reviewed by our Data Access Committee beforehand and granted by the archive only after its approval.”

Moreover, the module only covers the process of data sharing; Data Submitters are responsible for making sure that consent forms contain all other relevant information in order to be considered informative and transparent. For instance, we broadly define datasets within the GHGA Data Infrastructure as “human omics and related health data”, however consent forms should be clear on which specific kinds of data are collected and which specific kinds of data are shared for which secondary research as well as that the secondary research is framed appropriately and that relevant information on data security is provided.

The De-Identification Module

Omics and related health data should be de-identified before being broadly shared for scientific research. Accordingly, the Data Sharing Module explicitly speaks of the sharing of “pseudonymised data”. Here, we refer to the process of removing direct identifiers as

"pseudonymisation". Other terms (such as "coding") may be used as long as terms are used consistently throughout the consent form.[15] The De-Identification Module, available both in English (Table 3) and in German (Table 4), provides further information on the process of removing direct identifiers and the security status of de-identified data.

In order to protect your privacy, your data is pseudonymised before we share it with other researchers. That means, direct personal identifiers (such as your name and date of birth) are removed and replaced with a random series of letters and signs. Your pseudonymised data is no longer directly linked to you. A pseudonymisation key, which could be used to resolve the pseudonymisation and (re-)identify you, is kept separately and remains with the institution who carried out the pseudonymisation. The key will not be shared with others. Pseudonymised data are not anonymous. Certain kinds of data (e.g., genome data) could still be linked back to you, if combined with information about you available elsewhere (e.g., on genealogy databases). Due to the residual risk of (re-)identification from your pseudonymised data, we only make it available under controlled access conditions.

Table 3: De-Identification Module (English)

Um Ihre Privatsphäre zu schützen, werden Ihre Daten pseudonymisiert, bevor wir diese mit anderen Forschenden teilen. Das heißt, direkte persönliche Kennungen (wie ihr Name und Geburtsdatum) werden durch eine zufällige Folge von Buchstaben und Zahlen ersetzt. Ihre pseudonymisierten Daten sind nicht länger direkt Ihrer Person zugeordnet. Ein Pseudonymisierungs-Schlüssel, mit dem die Pseudonymisierung aufgelöst und die Daten (wieder) Ihnen zugeordnet werden könnten, wird separat aufbewahrt und bleibt bei der Institution, die die Pseudonymisierung durchgeführt hat. Der Schlüssel wird nicht mit anderen geteilt. Pseudonymisierte Daten sind nicht anonym. Manche Arten von Daten (bspw. Genomdaten) könnten immer noch ihrer Person zugeordnet werden, sollten diese mit anderen Informationen über Sie in Verbindung gebracht werden, die anderswo (bspw. in Ahnenforschungsdatenbanken) verfügbar sind. Aufgrund des Restrisikos einer (Re-)Identifizierung durch Ihre pseudonymisierten Daten, stellen wir diese nur unter kontrollierten Zugangsbedingungen zur Verfügung.

Table 4: De-Identification Module (German)

[15] Some might prefer the term "coding" in the context of genome data. This is so because "pseudonymised data" might be understood as data that allows (re-)identification only if one has the pseudonymisation key. However, in the case of genome data (re-)identification can be achieved by other means as well, e.g., through linking the data with information from other sources. When using the term "pseudonymised data", it is therefore important to describe the risks of (re-)identification clearly, including the risks resulting from data linkage.

Data Submitters may wish to state explicitly who performs the pseudonymisation (the Data Submitter themselves or a third party) and may wish to elaborate on the risks of (re-)identification. For instance, they may wish to explain in more detail which kinds of information about the Data Subject (e.g., on genetic predispositions) could be used to disadvantage the Data Subject, directly or indirectly, and which parties (e.g., insurance companies) might have an interest in obtaining such information.

Data Submitters may also wish to explain why it is important to keep the pseudonymisation key in the first place (e.g., for return of incidental findings). In any case, further information on the pseudonymisation process and responsibilities should be available on request.

The Controlled Access Module

Datasets within the GHGA Data Infrastructure and comparable scientific research data infrastructures are made available to researchers and research institutions under controlled access conditions. This is mentioned in the Data Sharing Module; but the consent form should elaborate clearly on how the controlled access process works. The Controlled Access Module, available both in English (Table 5) and in German (Table 6), provides further information on the controlled access process.

External researchers may access and use your archived, pseudonymised data only after a successful application. Researchers must first submit a request indicating which datasets they wish to use, how, for which purposes, and who would have access to the data. Each individual request is reviewed by us to make sure that it is in line with this consent. Only after our approval of a request does the archive grant the appropriate researchers access to your data. Researchers only have access to your pseudonymised data and do not receive any other information about you from us or the archive.

Table 5: Controlled Access Module (English)

Externe Forschende können Ihre archivierten, pseudonymisierten Daten nur nach erfolgreichem Antrag abrufen und nutzen. Forschende müssen zunächst einen Antrag stellen, der angibt, welche Datensätze sie wie und für welchen Zweck nutzen möchten, und wer Zugriff auf die Daten hätte. Jeder individuelle Antrag wird von uns geprüft, um sicherzustellen, dass Ihre Daten nur in Übereinstimmung mit dieser Einwilligung genutzt werden. Erst nach unserer Genehmigung eines Antrags gewährt das Datenarchiv den betreffenden Forschenden Zugang zu Ihren Daten. Forschende haben nur Zugriff auf Ihre pseudonymisierten Daten und erhalten von uns oder dem Archiv keine weiteren Informationen über Sie.

Table 6: Controlled Access Module (German)

Data Submitters should ensure that they can fulfil the requirement that external researchers "do not receive any other information about you from us or the archive". It should thus be clear to the Data Subject precisely what kinds of information are shared and what risks are associated with the sharing of that kind of information.

Data Submitters should state explicitly who reviews data access requests and should elaborate on the criteria used to arrive at data access decisions. Further information on the data access process should be available on request. If the Data Submitter appoints an external DAC, they should say clearly that it is the external DAC that reviews data access requests.

The Consent Options Module

It is important that the primary purpose of data use is distinguished clearly from secondary research purposes. Accordingly, consent forms should give Data Subjects a separate option to opt-in or opt-out of the sharing of their data for secondary research use, regardless of whether they consent to the primary purpose or not. The Consent Options Module, available both in English (Table 7) and in German (Table 8), allows the Data Subject to record their consent decision concerning downstream data sharing.

○ I agree that my pseudonymised data may be deposited in the German Human Genome-Phenome Archive (www.ghga.de) or other scientific data archives offering equivalent data protection standards and may be used for the named secondary research purposes within the limits set by this consent form. ○ I reject the archiving and use of my pseudonymised data for secondary research purposes as described in this consent form.

Table 7: Consent Options Module (English)

○ Ich stimme zu, dass meine pseudonymisierten Daten im Deutschen Humangenom-Phenomarchiv (www.ghga.de) oder in anderen wissenschaftlichen Datenarchiven mit gleichwertigen Datenschutzstandards gespeichert und für die genannten sekundären Forschungszwecke innerhalb der in diesem Einwilligungsdokument beschriebenen Grenzen verwendet werden dürfen. ○ Ich lehne die Archivierung und Nutzung meiner pseudonymisierten Daten für sekundäre Forschungszwecke, wie in diesem Einwilligungsdokument beschrieben, ab.

Table 8: Consent Options Module (German)

It is possible that the original consent form already asks Data Subjects to authorise the long-term storage and (re-)use of their data for secondary research purposes. If so, Data Submitters may prefer not to include a separate option to reject data sharing via the GHGA Data Infrastructure, but instead consider sharing via GHGA as part of the authorisation for the secondary research processing.

Final Evaluation of Consent Forms

After Module Integration

Consent forms that have been updated with the Modular Consent Toolkit require subsequent assessment to ensure that the consent text is consistent and coherent. The checklist below should help Data Submitters in performing the necessary steps to evaluate updated consent forms.

Please note that updated consent forms may require review by an ethics committee, even if the original consent form has already passed ethics review.

Checklist

The following checklist may be used to assess whether consent forms which have been updated with the Modular Consent Toolkit are consistent and coherent.

Among other things, updated consent forms should **not** contain any:

- Direct or indirect contradictions, e.g., about who is responsible for making data access decisions or who has access to data.
- Unnecessary repetitions, e.g., when explaining the data sharing and access process.
- Inconsistent language or wording, e.g., when describing de-identification as “pseudonymisation” or “coding”.
- Unclear references, e.g., it must be clear to the Data Subject who is meant when terms such as “we” and “us” are used. (It is reasonable for the Data Subject to assume that “we” refers to the researchers or the institution handing out the consent form to them, i.e., usually the Data Submitter.)
- Explicit or implicit ambiguities, e.g., regarding different infrastructures in which data may be stored or different purposes for which data may be used.
- Misleading structure, e.g., primary purposes of data collection should be described first before turning to the sharing of data for secondary research use.

Besides particular responsibilities highlighted in this document, in general, it is the responsibility of the Data Submitter to ensure that their consent form:

- Is compliant with data protection law, including clear elaboration on the rights and freedoms of Data Subjects, the responsibilities for their implementation, and naming of the relevant contacts for Data Subjects.
- Clearly elaborates on the aspects which can reasonably be assumed to be relevant to the consent decision, e.g., the (re-)identification risk inherent to pseudonymised data and data linkage, and the risk attached to the transfer of data to a data processor such as a data archive.

- Provides a lawful legal basis for data sharing. In this regard, Data Submitters should always consult their data protection officer as well as their legal department.
- Data Submitters should also contact their data protection officer for assistance in communicating informatively and transparently with Data Subjects when complying with their information obligations, e.g., on risk assessments and their results.

Conclusion

Outlook

It is expected that in the future, more and more genome and omics data will be generated as part of routine care, and it would be beneficial to make these datasets available for research. It is therefore important to GHGA to explore ways, together with the MII, to archive and use these routine data via a national omics archive.

Acknowledgements

This work has been done by the GHGA Consent Task Force as part of the GHGA project (NFDI 1/1), funded by the German Research Foundation (DFG) within the framework of the National Research Data Infrastructure (NFDI). We thank our funders as well as all co-authors and commentators for their valuable contributions.

Contact Information

If you have questions or comments, please feel free to contact the corresponding authors: Andreas Bruns (andreas.bruns@med.uni-heidelberg.de) or Eva Winkler (eva.winkler@med.uni-heidelberg.de).

License Type