










OPEN
ARTICLE

DS-PACK: Tool assembly for the end-to-end support of controlled access human data sharing

Pinar Alper^{1,2}  , Vilém Děd^{2,3} , Sascha Herzinger^{2,3}, Valentin Grouès^{2,3} , Sarah Peter^{2,3}, Jacek Lebioda^{1,2}, Linda Ebermann^{1,2}, Marina Popleteeva^{2,3}, Nene Djenaba Barry^{1,2}, Danielle Welter^{1,2}, Soumyabrata Ghosh^{2,3} , Regina Becker^{1,2}, Reinhard Schneider^{2,3} , Wei Gu^{1,2} , Christophe Trefois^{1,2} & Venkata Satagopam^{2,3}  

The EU General Data Protection Regulation (GDPR) requirements have prompted a shift from centralised controlled access genome-phenome archives to federated models for sharing sensitive human data. In a data-sharing federation, a central node facilitates data discovery; meanwhile, distributed nodes are responsible for handling data access requests, concluding agreements with data users and providing secure access to the data. Research institutions that want to become part of such federations often lack the resources to set up the required controlled access processes. The DS-PACK tool assembly is a reusable, open-source middleware solution that semi-automates controlled access processes end-to-end, from data submission to access. Data protection principles are engraved into all components of the DS-PACK assembly. DS-PACK centralises access control management and distributes access control enforcement with support for data access via cloud-based applications. DS-PACK is in production use at the ELIXIR Luxembourg data hosting platform, combined with an operational model including legal facilitation and data stewardship.

Introduction

Clinical and translational research relies on the use of biomedical data collected from human subjects, often called “human data”. Human data differs from other research data due to its sensitivity and personal nature. Collecting, handling and sharing human data requires preserving subject privacy and data confidentiality. The research community has developed the so-called “controlled access” model for data sharing to address these concerns¹. In this model, a researcher who wants to re-use human data from prior studies needs to make a formal access request to data controllers. The request is reviewed by a data access committee (DAC), composed typically of investigators responsible for the primary data/sample collection. The DAC review provides the necessary oversight and authorisation, ensuring the proposed reuse of data honours the use conditions placed by data donors and meets the necessary ethical and legal requirements. It is the responsibility of the researchers who produce data and would like to share it to ensure that the necessary operational procedures are in place so that the data is ingested into the controlled access realm, is findable through well-defined metadata, and is accessible via documented auditable processes.

Putting in place controlled access for one or more datasets is a complex and resource-intensive undertaking often performed by specialist intermediaries rather than researchers themselves². The most common way for researchers to provide a controlled access layer over research data is either by depositing the data into centralised human genome/phenome archives^{3,4}, or by handing data over to a “data support” team so that that team becomes the contact point for data access requests and handles the downstream processes. In both models, these intermediaries perform some or all of data discovery, access request management, data storage and delivery functions.

In recent years, several developments have diminished the choice of centralised repositories. First, the European General Data Protection Regulation (GDPR)⁵ has restricted the cross-border movement of data as it

¹Luxembourg National Data Service, PNED GIE, Esch-sur-Alzette, L-4362, Luxembourg. ²ELIXIR Luxembourg, Belvaux, Luxembourg. ³Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, L-4367, Luxembourg. [✉]e-mail: pinar.alper@elixir-luxembourg.org; venkata.satagopam@uni.lu

requires additional safeguards during data collection and sharing^{6,7}. Researchers now need to consider complex ethico-legal aspects before depositing data to archives outside the country of data collection and/or the EU, as obtaining necessary safeguards may be too time-consuming or, in some cases, impossible. To accommodate GDPR requirements, repositories have proposed “federated” data sharing models⁸. In a repository federation, the data discovery function remains centralised; meanwhile, data storage and delivery are handled by distributed, often national, nodes. The emerging federated approach succeeds in keeping data within national borders; however, it currently has two functional gaps. First, managing access requests is left to the nodes for them to coordinate – among requestors, providers and access committees – the review of the requests and the conclusion of respective data use agreements.

Second is the evolution in modalities of data access and analysis. The traditional mode of data access has been file downloads by authorised requestors. We are now seeing – partly due to the GDPR – an increase in the adoption of the so-called “compute to data” approach, where being granted access to a dataset means being able to access and run analysis in a cloud environment that contains the dataset rather than being able to download the data. Current federation implementations support data access via file downloads, for clinical and translational data in particular; this approach prevents the incorporation of various data-sharing cloud platforms into federations⁹.

The landscape changes initiated by the GDPR have become the impetus behind institutional efforts to establish local data support teams and controlled-access processes. The challenge that awaits these teams is (1) a lack of software tools that would support implementing the controlled access process, (2) a lack of formalised roles for the ownership and improvement of the process¹⁰ and (3) the necessity to address GDPR accountability requirements¹¹, such as audit trails and documentation, without dragging the whole process.

As a member of ELIXIR¹², a pan-European life science information infrastructure, ELIXIR Luxembourg (<https://elixir-luxembourg.org>) is the data support partner in several European projects. Over the years, we have had to overcome the said challenges. As of today, ELIXIR Luxembourg hosts a data catalog¹³ and provides controlled access to diverse types of human datasets in various modalities, either via file transfer or in cloud environments and applications, whereby data is brought to life through curation, integration and analytics. The solution blueprint that underlies our controlled-access setup is the Data Stewardship Provenance and Compliance Kit (DS-PACK). DS-PACK is an assembly of open-source tools that:

- acts as middleware translating the information from the data access request management system to commonly used authentication and authorisation protocols, thereby allowing the sharing of data via any hosting platform supporting these protocols;
- solves a problem inherent in distributed and federated data sharing, which is the centralisation of access control over distributed, heterogeneous hosting platforms;
- supports the sharing of diverse biomedical data types and data delivery in different modalities;
- semi-automates the controlled access processes by outlining clear roles for data support teams, which can relieve pressure from DACs and help scale up data sharing;

In this paper, we provide an overview of DS-PACK and illustrate how it has been put to production use within ELIXIR Luxembourg’s data hosting service. We outline the three core contributions of DS-PACK as (1) a standards-compliant access control pattern implementation for research data platforms, (2) an open-source and reusable component assembly, and (3) an implementation of the data protection by design and by default (DPbDD) principle of the GDPR. We present the method adopted to build DS-PACK. Finally, we review related work and discuss how DS-PACK relates to existing approaches and discuss our future work.

Results

Centralised access control for distributed, heterogeneous data hosting platforms. When data is shared over distributed platforms, an inherent problem is managing access control lists (ACLs) for each platform. List management becomes a bottleneck as the number of platforms and granted accesses increase. DS-PACK adopts an existing information security pattern to address this issue, where the ACL is maintained centrally and necessary information is propagated to hosting platforms during data access.

Figure 1 illustrates ACL management in DS-PACK, and the next section describes its software components. We use the Data Information System (DAISY)¹¹ to store and maintain the central ACL. Access information is a triple, which is the combination of (1) the user having access, (2) the persistent identifier (PID) for the dataset given access, and (3) the provenance of the access record, including time of creation and the record’s source. Accesses are created following the access request review.

DS-PACK uses the Open ID Connect (OIDC) standard¹⁴ to authenticate the user and it uses JSON Web Tokens (JWT)¹⁵ to transport authenticated user identity and permissions. We use Keycloak (<https://www.keycloak.org>) as the identity and single sign on platform. Keycloak supports external identity providers (IdP); user authentication can be performed in a federated manner, allowing users to log in with an existing account in any other trusted IdP that supports the OIDC standard, e.g. user’s home institution, ORCID or Life Science AAI Login (<https://elixir-europe.org/platforms/compute/aai>). Any information the IdP provides can also be collected and used for internal user management, e.g. utilising user affiliation information in access request handling and review.

DS-PACK includes mappers - one per hosting platform - that pack authorisation information into JWT tokens so the recipient platform can consume it without code changes. Transporting authorisation information in user tokens is an efficient and adopted pattern in information security. Such a solution allows authorisation to be propagated only to the platforms the user accesses without custom propagation logic.

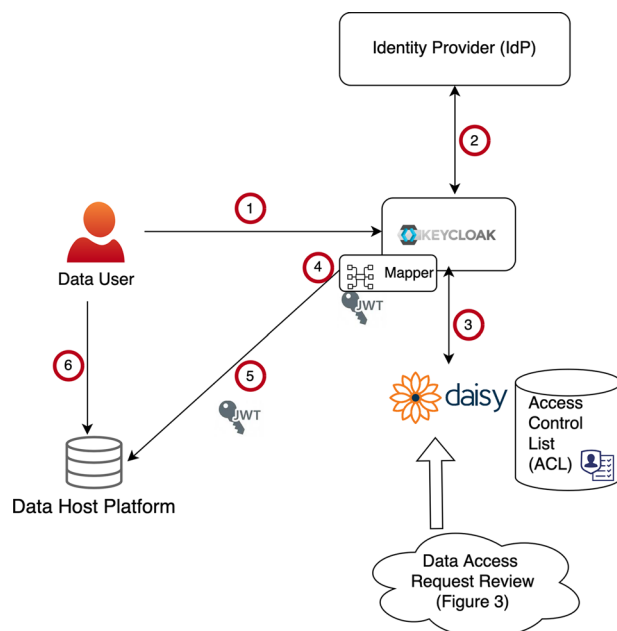


Fig. 1 Authentication and access control in DS-PACK. (1) User is directed by the data host platform to Keycloak for authentication. (2) User is authenticated via their associated identity provider. (3) Keycloak pulls the user's permissions from the access control list in DAISY. (4) Using mappers for the data host platform the permission information is packed into JWT tokens. (5) The user identity and permissions are presented to the data host platform in JWT tokens. (6) If permitted, the user can access data on the platform.

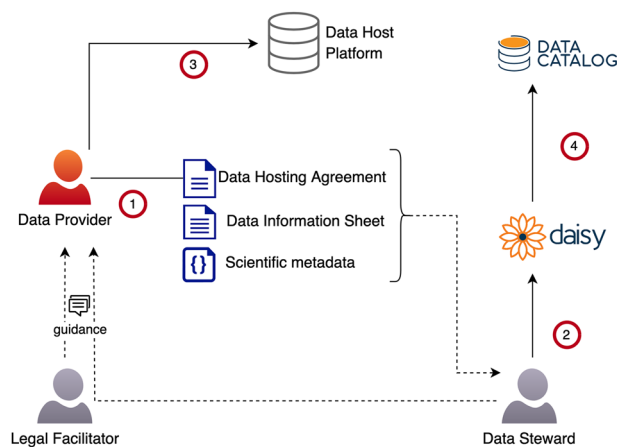


Fig. 2 ELIXIR Luxembourg data submission process based on DS-PACK. (1) The data provider customises the data hosting agreement template to suit their requirements with guidance from the legal facilitator, provides access policy and ELSI, and GDPR metadata by filling out the Data Information Sheet (DISH) with guidance from the data steward, and provides scientific metadata for the dataset in DATS JSON format. (2) The data steward imports the DISH and the DATS JSON into DAISY to create the corresponding dataset record(s); the data hosting agreement is attached. (3) The data provider makes the data available on a host platform maintained either by the provider or by ELIXIR Luxembourg. (4) The data steward finalises the dataset record by adding information on data (host) locations, and publishes the dataset; as a result, the dataset gets a persistent identifier and becomes visible in the Data Catalog.

Open-source tool assembly for the end-to-end support of the controlled access process. The DS-PACK aims to provide automation support for the controlled access process. We achieve this with an integrated assembly of software tools, metadata collection and agreement templates; we also provide a recommended operational model and identify roles for intermediaries similar to ELIXIR Luxembourg. DS-PACK is based entirely on open-source tools and templates, which implement standards where applicable. Component descriptions are given in following subsections, and the assembly's operation for ELIXIR Luxembourg's data submission and access process are given in Figs. 2, 3 (user authentication is omitted from these figures, as it is discussed in the previous section).

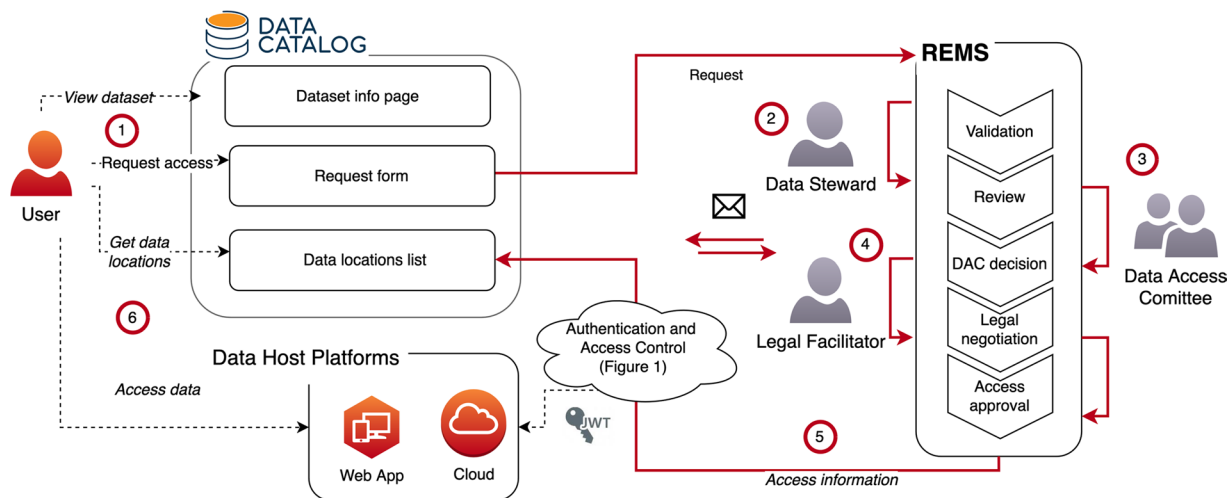


Fig. 3 ELIXIR Luxembourg data access process based on DS-PACK. (1) The user locates the dataset in the catalogue and initiates an access request filling out the form associated with the dataset. (2) The data steward reviews the request and forwards it to the DAC. Additional information can be received from user via email. (3) The DAC reviews the request and makes a decision. During the review, the DAC may ask the data steward for further information. (4) Following a positive decision, the user customises the data use agreement template with guidance provided by the legal facilitator. (5) When the data use agreement is signed, the decision result is propagated from REMS to the access-control master in DAISY. (6) The user accesses the data in the data host platform. For each access attempt, Keycloak will consult the access control list in DAISY and present the user's permissions to the host platform in their access token.

The DS-PACK operational model provides end-to-end data submission and access support, including legal facilitation and data stewardship. These two activities, which are further described in this section, make up the two main manual parts of the process. ELIXIR Luxembourg concludes agreements with providers and users separately, removing the need for these parties to enter legal processes in a peer-to-peer fashion. We provide templates for data hosting and data use agreements, which can be used as-is or with minimal customisations by legal facilitators. The combination of templates and facilitators brings efficiency to the legal processes and ensures that the necessary provisions concerning protecting sensitive human data are in place.

The data providers submit ELSI metadata for the datasets shared under controlled access, including use conditions, in data information sheets (DISH). They also provide additional structured metadata about the content of the datasets and the modalities of the studies from which they were generated. Specifically, metadata describes (1) the datasets that are shared under controlled-access, (2) the data sources, specifically the cohort studies in which the clinical and translational datasets were collected, (3) the data providers, namely the research project and principal investigators that are making the data available for re-use. Details of the schemas and ontologies utilised for metadata are described in the next section under the "Data Catalog".

Data stewards import ELSI and scientific metadata into DAISY to obtain draft metadata records. ELSI metadata is further curated and verified by reviewing the provider's access policy and access agreement documents. Once finalised, data stewards publish the metadata in the data catalog, making datasets discoverable by prospective data users.

Data stewards support the access process by validating the identity and affiliation of the requestors and by an initial assessment of their eligibility to access data based on defined use conditions. Requestor registration and identity checks are typical in secondary use of human genome/phenome data and have been formalised as "registered-access"¹⁶. The identity checks performed by data stewards are intended to offload such duties from the data access committee so that the committee can focus checking whether the proposed secondary use of data in the requestor's research project is inline with the purposes for which the data is collected.

Data stewards also mediate the communication between requestors, the data access committee (DAC) and the legal facilitator, maintaining frequently asked questions on datasets and collecting from the requestor any further information required by the DAC.

DS-PACK Components: software tools. **Data Information System (DAISY)**¹¹ is a dataset registry and documentation tool meeting GDPR accountability requirements for biomedical research projects. DAISY is targeted for the use of data stewardship teams and is central to the operation of DS-PACK, acting as ELSI metadata and access control master.

Keycloak (<https://www.keycloak.org>) is a platform for single sign-on with identity and authorisation management. It implements the Open ID Connect (OIDC) protocol¹⁴, which uses signed and verifiable JSON Web Tokens¹⁵ to transmit user authentication and authorisation information to clients.

Data Catalog¹³ is a dataset information index and search tool based on the DAta Tag Suite (DATS) schema¹⁷, which is fully interoperable with the W3C Data Catalog Vocabulary (DCAT)¹⁸. Discipline-specific metadata is then added to catalog records in the form of key-value pairs, thanks to the flexibility of the DATS model. A

	DISH	DAISY	Keycloak	Data Catalog	REMS	Data Host Platform	Agreements
Transparency, Lawfulness, Fairness	x	x			x		x
Purpose limitation	x	x		x	x		x
Data minimisation	x	x	x		x		x
Storage limitation		x				x	x
Confidentiality and integrity		x	x			x	x
Accountability		x	x	x	x	x	x

Table 1. GDPR data protection principles supported by the DS-PACK components.

range of ontologies are used in the Data Catalog, including but not limited to the Data Use Ontology (DUO)¹⁹, the SemanticScience Integrated Ontology (SIO)²⁰, the National Cancer Institute's Thesaurus (NCIt)²¹, the MONDO Disease Ontology²², the EDAM Bioinformatics Ontology²³, the Units Ontology (UO)²⁴ and the CHEBI Chemistry Ontology²⁵.

Resource Entitlement Management System (REMS)²⁶ is a tool for creating and executing workflows assessing data access requests. Applicants can utilise their federated user IDs to access REMS, complete the data access application, and acknowledge the use conditions associated with the resource. Once the application has been submitted, REMS assigns it to the workflow handler, e.g. a data steward or a data access committee member, for review and approval. Additionally, REMS can generate reports detailing the status of the applications and the data access rights that have been granted.

Data Host Platform is any platform that implements the OIDC standard and consumes authorisation information from JWT tokens. Currently, we have integrated the ADA Platform (<https://ada-discovery.github.io>) into DS-PACK, and we are working on integrating RedCAP²⁷. Host platforms have to implement de-serialization and resolution to platform-specific access permissions. Timely exchange of authentication and authorisation information with Keycloak, i.e. upon each login and token expiry, is also a responsibility of the platform.

DS-PACK Components: templates. **Data information sheet (DISH)**²⁸ is a metadata collection instrument, organised as a questionnaire spreadsheet, to capture information on Ethical Legal and Societal Issues (ELSI) concerning the re-use of data. We refer to this information as “ELSI Metadata”²⁹, primarily composed of data use conditions stemming from consent clauses, data provider policies and access agreement conditions.

Data hosting agreement³⁰ is a legal document signed between the data host and the data provider(s) that contains the provisions for long-term data hosting, GDPR-compliant data access to third parties, as well as additional data management services, where needed, such as data curation or (re)pseudonymisation. The provider signs the data hosting agreement and, in the annexe, provides the DISH. The Data Access Policy is built upon the information provided by the Data Provider through the DISH.

Data use agreement³¹ is a legal document signed between ELIXIR-LU and the data user's institution that contains the provisions for the GDPR-compliant use of data, including use purpose limitations, storage durations, non-transferability, right of data subjects and information security safeguards.

Data user responsibilities acknowledgement³² is a declaration which accompanies the data use agreement; it is signed by each data user that is listed as a prospective data accessor in the access request form; users confirm they will comply with the data use conditions and general good scientific practice for data handling.

An implementation of data protection by design and by default. Data protection by design and by default (DPbDD) is a provision of the GDPR⁵ requiring data controllers to think ahead on data protection and incorporate appropriate technical and organisational measures into the design of data processing systems. DS-PACK assembly follows DPbDD through its four key features: ELSI Metadata, DAC Review, Centralised ACL management and model agreements. We discuss, next, how each DS-PACK component supports GDPR's data protection principles through these features. A mapping of components to principles is also given in Table 1.

Transparency, lawfulness, fairness. ELSI metadata captured in DISH and recorded in DAISY requires the data providers to declare the legal bases for data collection, sharing and secondary use, respectively. Using REMS, the DAC verifies whether the proposed use conforms to data use conditions originating from informed consent, contributing to transparency and fairness of secondary use of personal data. The data agreements, both the data hosting agreement and the data use agreement, oblige the data host/repository, the data provider and the data user to comply with Article 5 of the GDPR, which lays out the core data protection principles listed in Table 1.

Purpose limitation. A crucial part of ELSI metadata is data use conditions, which outline the permitted and prohibited uses of data and the obligations that shall be fulfilled before access. Respecting conditions that descend from participant informed consent is a primary responsibility for data sharing under the GDPR^{33,34}. Figure 4 displays use conditions originating from informed consent, such as disease-specific research use, and conditions commonly required by data providers, such as ethics review or collaboration requirements.

Publishing use conditions in the data catalog ensures that prospective data users are informed when requesting access. The DAC review of the proposed data use then verifies conformance to data use conditions. The data hosting agreement holds providers liable for providing metadata that allows GDPR-compliant processing of the data by the host/repository and the data users; meanwhile, the data use agreement restricts the data user's use of the data to the DAC-approved purpose.

Request access to 'PRECISESADS IMI-JU GA 115565 SUSTAINABILITY STUDIES'

Please provide a brief description of your intended data use: *

Data Access Process acknowledgment *

 I confirm I have read and understood the Elixir Hosting Data Access Process (<https://elixir-luxembourg.org/services/catalog/data-hosting/data-user/>) *

Use Restrictions

All use restrictions must be accepted to be able to request access to this dataset

 Constrained permissions
 Use is limited to research in the field of systemic auto immune diseases (SADS) (Disease-specific Research and Clinical Care.) *

 Use is limited to non-commercial scientific research project and publication purposes. Such non-commercial use by a commercial entity is permitted. (Not-for-profit Use Only) *
 Obligations
 Acknowledgement is required: This work has received support from the EU EFPIA Innovative Medicines Initiative, Joint Undertaking PRECISESADS grant no 115565. Where reasonable consider if partners of PRECISESADS may be co-authors. (Publication Required) *

 The Sustainability Committee established by the Data Providers must be informed of any publication using PRECISESADS data. As a result, the user institution must inform ELIXIR-LU of any publication concerning these data, regardless of the frequency (Publication Required) *

SEND

Explanation of use restriction code can be found on the GA4GH website

* mandatory fields

Fig. 4 An example access request form. The user is presented with all data use conditions and the required data access agreement. Best practices for studies collecting human data recommend consent with “broadly described research purposes with ongoing updates for participants” as well as allowing “participants to retain control”⁴⁴. The “Disease-Specific Research”^{19,45} use condition illustrated in this screenshot is common in controlled access human data sharing and it emerges from the so-called “Tiered-Consent” model⁴⁶. This model allows study participants to retain more control over their data by giving them the option to consent to data uses in particular research categories or settings. Compared to broad consent, the tiered model can be seen as a compromise allowing participants to opt in to share their data, which otherwise would be confined to the primary study and not shared.

Data minimisation. providers are asked to confirm that the data is pseudonymous and contains no standard personal data attributes as part of the ELSI metadata declaration. The DAC also reviews whether the proposed research requires all variables of the data or whether a data subset needs to be requested. Accesses granted are contractually time-bound for a default period of one year; after this, researchers need to renew their access request and the data use agreement.

Storage limitation. For datasets hosted in the repository, DAISY acts as the central register of data storage locations and data hosting platform endpoint URLs. The data hosting agreement mandates limited durations of data storage by data users, which are tracked in DAISY. Data stewards get notified of datasets nearing the end of their storage period. In addition, by allowing cloud data-sharing platforms that access without data download, the DS-PACK eliminates the need for storage limits tracking with the users of such platforms. To support research reproducibility, the data hosting agreement foresees long data retention periods for the host repository. The agreement is concluded typically for a minimum of 10 years, and it contains provisions for an additional data retention period of 15 years upon agreement termination.

Confidentiality and integrity. Ensuring confidentiality is the primary goal of DS-PACK. Researchers can access only those datasets for which they have been granted permission. Accesses are centrally managed and globally enforced, a design pattern implemented with DAISY, Keycloak and data platform integration. Upon a positive DAC response, access expiration is set by the workflow handler in REMS and propagated to DAISY. Upon completion, the access requires an application for renewal and a new DAC review. DS-PACK adopts conservative login policies; user tokens are short-lived, and logins to a platform are only allowed when at least one permitted dataset exists on that platform. Confidentiality obligations are standard components of all agreements. In particular the data use agreement outlines responsibilities for the user’s host organisation in cases of misconduct such as attempting to re-identify data subjects or compromise subject identity. Data stewards can manually revoke access in case of misconduct or when the users leave the institution before the end of the access expiry period.

DS-PACK handles and grants access per named user and does not support user groups. Individual data hosting platforms handle data confidentiality during storage, analysis and transfer. There is no technical barrier to having user groups and assign permissions to those. We observe, however, that having groups adds a level of indirection to the representation of ACLs, as user permission is no longer an explicit record but would need to be deduced from group membership. This indirection bears potential for obfuscating the audit trail of accesses obtained and lost by users. In addition, our agreements list named users and we see value in user participation in the access request process to raise awareness of data use conditions and user responsibilities.

Accountability. DS-PACK addresses the accountability requirement by documenting sensitive datasets, accesses and the provenance of access decisions. DAISY is the central register of sensitive data and the logbook of activities concerning data, thereby implementing the “Register Of Processing Activities” (ROPA) outlined in the GDPR. The entire process, from access request to DAC review, the population of ACL lists and the enforcement of access are automatically logged. Logs are easily correlatable by user names and dataset persistent identifiers (PID). The audit trail for the DAC process is recorded in REMS, and individual accesses are logged in Keycloak and the hosting platforms. Any manual updates to ACL lists in DAISY such revocations are logged.

Finally, we have performed a Data Protection Impact Assessment (DPIA) for the ELIXIR Luxembourg data hosting setup based on DS-PACK. The DPIA outlines privacy risks and mitigations, thereby providing evidence of the decision-making behind the DS-PACK design. Our model agreements ensure that involved parties follow all principles and clearly define their accountability. There are differences in GDPR interpretation by the national data protection authorities and consequently differences in national implementations. Our agreement templates are GDPR-observant and general-purpose. In case they are to be used in countries with distinct national GDPR provisions, then those would need to be reflected to the agreements.

Discussion

Related work. The US-based dbGaP⁴ is one of the oldest deposition databases for subject-level genome/phenome data; as such, it has set a precedent for the controlled access model and various other data sharing initiatives³⁵. dbGaP requires data submitters to delineate all data use conditions as data use limitation tags on datasets. To streamline data access, dbGaP concludes an online Data Use Certification Agreement digitally confirmed by the data requestor and a representative from their institution. Data provision is through time-limited file downloads, and dbGaP notifies data users to delete local copies after one year or to renew their access.

EGA³ is a controlled access repository of the European Molecular Biology Laboratory and the Center for Genomic Regulation. EGA provides partial support for the controlled access process focusing on dataset discovery and provision, leaving out the DAC orchestration and the agreement facilitation. EGA users are provided with the contact details of DACs associated with datasets from whom they can request access. As guidance to DACs, an access agreement template is provided; each DAC, however, is responsible for drafting their agreement template and concluding individual agreements with requestors.

The EGA is developing a “federated” model (FEGA), where data discovery occurs via the central repository and the data hosting and provision is done by (local) nodes⁶. Like the dbGaP, EGA and FEGA data provision is also based on file downloads. A federated sharing model is arguably more complex than a centralised one. Despite its complexity, the federated model has emerged as a response to the GDPR. Particularly GDPR-required safeguards for cross-border data transfer and a lack of clarity and shared interpretation at the EU level³⁶ make it difficult for researchers to deposit data to central repositories. E.g. data collected from a Luxembourgish cohort without consent provisions for sharing outside the country cannot be transferred to centralised repositories. An advantage of federations, which keep the data close to source is that each node can continue to complete and curate the data without the need to transfer different releases to a central repository and instead transfer its metadata only.

GA4GH Passports² is an open standard that focuses on transporting a user’s access rights, called visas, along with the user’s identity, called a Passport. The standard outlines visa types, representing the typical information that must be transported when researchers access cloud environments holding open or controlled access genomics and health datasets. Examples of visas are a researcher’s institutional affiliation, role, other linked identities and finally, the datasets to which they have been granted access. The Passports standard is based on the OIDC standard; as such, it uses digitally signed JWT tokens that ensure the identity and access information in tokens are authentic and verifiable. GA4GH Passports is a recent standard for which few demonstrators have already implemented, including the test implementations for the dbGaP and the EGA repositories.

Sage Synapse³⁷ and the Open Science Framework³⁸ are two research collaboration and results sharing platforms that promote Open Science practices. For sensitive human data, these platforms offer the “controlled access” option. The Synapse platform requires data contributors to designate data use conditions. The platform facilitates an access request process directly between the data requestor and the contributor not utilising DACs or data sure agreements. The OSF, on the other hand, achieves controlled access by directing its users to designated repositories.

DS-PACK present and future. DS-PACK has been in production use at ELIXIR Luxembourg since 2021 to deliver our node’s data hosting service. We have coupled DS-PACK with an operational model, where legal facilitators and data stewards act as intermediaries by tailoring agreements, facilitating DAC-user communication, data request validation and ensuring ELSI metadata precision. In this regard, our solution represents a middle ground between those repositories with highly streamlined access procedures and fixed agreements, such as the dbGaP, and others that leave out support for legal facilitation and DAC orchestration such as the EGA. DS-PACK operational process involves human elements, which raises the question of scalability. Centralised repositories such as the dbGaP and the EGA have succeeded in vertically scaling data sharing with large numbers of datasets

and access requests. Now the emerging federated models call for a horizontal scaling of data sharing. The goal of DS-PACK is to be part of that horizontal scaling by empowering institutional data support teams and to establish data sharing nodes in federations.

Data support teams are lean; they undertake diverse responsibilities with limited human resources. Practical duties of ELIXIR Luxembourg data stewards involve responding to inquiries from data submitters and requestors, facilitating communication, curating ELSI metadata, and ACL list management. List management is not as time-consuming as other tasks, but it requires high accuracy in implementation otherwise leading to non-compliance. The automation-support that the DS-PACK brings both time savings, as in the case of REMS facilitating communication of various parties and increased accuracy as in the case of centralized ACL management.

Any system that supports OIDC-based authentication and authorization, can be connected as a data host platform to DS-PACK. The effort required for integration ranges within a few days of development or configuration work. A natural next step for us is to implement support for the GA4GH Passports specification and enable our “mappers” to generate Passport-compliant tokens from ACL lists. The use case for Passports emerges for data users navigating environments with different trust levels, such as when the user needs to access data in a cross-organisational research data cloud environment.

Throughout the various multi-party clinical and translational studies we were involved, we observed that data sharing starts as early as the data collection, and the audience with which the data is shared changes over time. During the study, the data is shared within the consortium for primary use, and it is subject to the study's ethico-legal framework, a model referred to as “clique sharing”³⁹. Upon study completion, the data is shared with the broader research community for secondary use, which may require different legal provisions. The DS-PACK operational model allows us to deploy the same tool assembly for both primary and secondary use, and we will continue to deploy it for various consortium studies. The loose coupling of the DS-PACK components allow us to deploy the assembly partially depending on projects' requirements.

The DS-PACK design emphasises the collection and curation of ELSI metadata by data stewards. Metadata meets the documentation requirements of the GDPR as it contains all attributes identified by Article 30, and more. This information provides a baseline to build the necessary technical and operational safeguards towards GDPR principles. Our way of modelling use conditions allows us to specify in a more granular, therefore more precise manner which data uses are permitted and prohibited; we can also incorporate terms from different vocabularies such as the Data Use Ontology (DUO)¹⁹. To facilitate indexing of our data catalog by other thematic catalogs, we are planning to expose our DATS-based catalog content in DCAT form.

ELIXIR Luxembourg will soon join one or more controlled access data sharing federations at the EU level^{8,40,41}. The DS-PACK development constituted our groundwork as middleware allowing us to establish our node repositories and operations in different federations. DS-PACK assembly is openly available for the use of research data support teams establishing thematic and/or institutional repositories or nodes within repository federations. All component software licenses allow commercial use; only the Data Catalog and DAISY require derivative software to stay open sourced. We are building easier containerised deployments and improving our documentation to assist adopters.

Method

ELIXIR Luxembourg already had a process to support controlled access for project consortia. Our process was largely manual, using email communication and document-based forms. DS-PACK development commenced around the time the GDPR came into effect. We started by mapping out the existing process and identifying process gaps and priorities for automation; we also identified opportunities for metadata standardisation. In parallel, an ELSI expert at our node translated the GDPR principles and other relevant legal requirements into prospective technical and operational measures that could be adopted.

Before the implementation, we reviewed existing open-source tools to assess whether they could be used in our solution. The implementation of DS-PACK was iterative and incremental; automation was gradually introduced. In the first phase, we deployed DAISY to build our internal inventory of sensitive datasets, including ELSI metadata. Next, we introduced the Data Catalog into the assembly, exposing dataset metadata in a standard and findable manner. Finally, we added REMS, Keycloak and hosting platforms with mappers to automate DAC review orchestration and enforce access decisions.

DAISY utilised GA4GH's Consent Codes, which are atomic data use terms representing common secondary use conditions for research data. We extended DAISY to store use conditions as triples, combining a use term with a use rule denoting whether the cited use term denotes permission, a prohibition or an obligation⁴². We added the ability to plug external PID generation services, e.g. DOIs, to support other PID types in addition to DAISY's default internal accessioning scheme. Upon dataset publishing, the user must select the publishing target (catalog) and an existing REMS form. To support ACL management, we extended the DAISY information model and integrations. DAISY already provided a data logbook allowing data controllers to record data lifecycle events, including accesses. We refined the access schema to accommodate the ACL triple.

We connected the Data Catalog to DAISY's REST endpoint to pull dataset descriptions, ELSI and discipline-specific metadata in DATS JSON format. This connection allowed the two systems to have synched information in near-real-time. We also extended the Data Catalog to act as a front end to REMS for the creation of access requests for selected datasets. The DAC review is triggered from the corresponding user form added to the data catalog interface. REMS is used as-is in DS-PACK, not requiring custom extensions. We configured the REMS to notify DAISY in case of granted accesses.

As Keycloak implements the OIDC standard, most of the custom work was limited to creating mappers for the JWT tokens that carry user permissions. For client applications hosting multiple datasets, the client must be capable of reading claims inside the JWT tokens and mapping them to certain local permissions. While this

is technically relatively trivial, many applications added OIDC only rudimentarily to their historically already existing authentication solution. Fortunately, OIDC is growing in popularity, so adoption and familiarity with the protocol are improving, and with them, the level of integration. In simple situations where one application hosts one dataset, the only requirement for the integration into our access pipeline is the support of OIDC for authentication. The JWT tokens we use are signed, and thus immutable, and sent over an encrypted connection (HTTPS), and therefore not readable by anyone but source and target. We have, however, not encrypted the token content as the content is not sensitive in-of-itself and therefore its encryption would not add any value to our solution.

To facilitate ACL lookup from DAISY, we developed a plugin that extends the Keycloak authentication flow with a call to DAISY's REST API when a given user logs in. By not relying on scheduled synchronisations but extracting this access information on-demand, we can ensure that the information is up-to-date when served to the client applications. We also added functionality to Keycloak that allows us to control access to an application at the authentication stage based on self-defined policies. In other words, we can prohibit users from logging into an application if they do not have a certain dataset access record. This allows us to integrate applications into our access control pipeline, even if the application only supports central authentication via OIDC but handles authorisation locally.

Data availability

The ELIXIR Luxembourg data catalog(<https://datacatalog.elixir-luxembourg.org>) currently lists 261 datasets; 27 of those are hosted in platforms managed by our node. Human datasets that are not of an anonymous nature fall under the GDPR, and therefore, they are available under a controlled-access regime such as the “PRECISESADS IMI-JU GA 115565 Sustainability Studies” dataset⁴³, which is used for the testing of the DS-PACK assembly. The process for requesting a controlled access dataset can be initiated by clicking the “Request data access” button on the dataset information page in the catalog.

Anonymous human datasets or, other, non-human data is available under open access via direct download by clicking the “Access data” button on the dataset information page.

Code availability

Source code for the components of the DS-PACK assembly can be found in the following repositories. Among these, ELIXIR Luxembourg is the main developer of DAISY, Data Catalog and ADA systems and the Keycloak DAISY synch plugin.

- Data Information System DAISY, GNU Affero General Public License (AGPL 3.0), <https://github.com/elixir-luxembourg/daisy>
- Data Catalog, GNU Affero General Public License (AGPL 3.0) <https://github.com/FAIRplus/imi-data-catalogue>
- Keycloak, Apache 2.0 License, <https://github.com/keycloak/keycloak>
- Keycloak - DAISY synch plugin, Apache 2.0 License, <https://gitlab.lcsb.uni.lu/keycloak-projects/keycloak-providers>
- Resource Entitlement Management System (REMS), MIT License, <https://github.com/CSCfi/remS>
- ADA Discovery Analytics (ADA), Apache 2.0 License, <https://github.com/ada-discovery>

The source code to connect DS-PACK components is located within each component. We are working on a containerised demo of DS-PACK execution, which will be made available under the ELIXIR Luxembourg GitHub repository. <https://github.com/elixir-luxembourg/>

Received: 3 October 2023; Accepted: 29 April 2024;

Published online: 15 May 2024

References

1. Resnik, D. B. Genomic research data: open vs. restricted access. *IRB: Ethics & Human Research* **32**, 1 <https://link.gale.com/apps/doc/A239462724/AONE?u=google scholar&sid=bookmark-AONE&xid=5d539e1d> (2010).
2. Voisin, C. *et al.* GA4GH Passport standard for digital identity and access permissions. *Cell Genomics* **1**, 100030, <https://doi.org/10.1016/j.xgen.2021.100030> (2021).
3. Freeberg, M. A. *et al.* The European Genome-phenome Archive in 2021. *Nucleic Acids Research* **50**, D980–D987, <https://doi.org/10.1093/nar/gkab1059> (2022).
4. Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Research* **42**, D975–D979, <https://doi.org/10.1093/nar/gkt1211> (2014).
5. European Parliament & Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. <https://data.europa.eu/eli/reg/2016/679/oj> (2016).
6. Becker, R., Thorogood, A., Bovenberg, J., Mitchell, C. & Hall, A. Applying GDPR roles and responsibilities to scientific data sharing. *International Data Privacy Law* **12**, 207–219, <https://doi.org/10.1093/idpl/ipac011> (2022).
7. Lawlor, R. T. The impact of GDPR on data sharing for European cancer research. *The Lancet Oncology* **24**, 6–8, [https://doi.org/10.1016/S1470-2045\(22\)00653-2](https://doi.org/10.1016/S1470-2045(22)00653-2) (2023).
8. Freeberg, M. & Curwin, A. Federated EGA Updates in 2022. *F1000 Research* <https://doi.org/10.7490/F1000RESEARCH.1118988.1>. Publisher: F1000 Research Limited (2022).
9. Athey, B. D., Braxenthaler, M., Haas, M. & Guo, Y. transSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* **2013**, 6–8 (2013).
10. Boeckhout, M., Zielhuis, G. A. & Bredenoord, A. L. The FAIR guiding principles for data stewardship: fair enough? *European Journal of Human Genetics* **26**, 931–936, <https://doi.org/10.1038/s41431-018-0160-0> (2018).
11. Becker, R. *et al.* DAISY: A Data Information System for accountability under the General Data Protection Regulation. *GigaScience* **8**, giz140, <https://doi.org/10.1093/gigascience/giz140> (2019).

12. Harrow, J. *et al.* ELIXIR: providing a sustainable infrastructure for life science data at European scale. *Bioinformatics* **37**, 2506–2511, <https://doi.org/10.1093/bioinformatics/btab481> (2021).
13. Welter, D. *et al.* The Translational Data Catalog - discoverable biomedical datasets. *Scientific Data* **10**, 470, <https://doi.org/10.1038/s41597-023-02258-0> (2023).
14. Sakimura, N., Bradley, J., Jones, M., de Medeiros, B. & Mortimore, C. *Openid connect core 1.0 incorporating errata set 1*. https://openid.net/specs/openid-connect-core-1_0.html (2014).
15. Jones, M. B., Bradley, J. & Sakimura, N. JSON Web Token (JWT). RFC 7519 <https://www.rfc-editor.org/info/rfc7519>, 10.17487/RFC7519 (2015).
16. Dyke, S. O. M. *et al.* Registered access: authorizing data access. *European Journal of Human Genetics* **26**, 1721–1731, <https://doi.org/10.1038/s41431-018-0219-y> (2018).
17. Sansone, S.-A. *et al.* DATS, the data tag suite to enable discoverability of datasets. *Scientific Data* **4**, 170059, <https://doi.org/10.1038/sdata.2017.59> (2017).
18. Albertoni, R. *et al.* The W3C Data Catalog Vocabulary, Version 2: Rationale, Design Principles, and Uptake (2023).
19. Lawson, J. *et al.* The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genomics* **1**, 100028, <https://doi.org/10.1016/j.xgen.2021.100028> (2021).
20. Dumontier, M. *et al.* The SemanticScience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics* **5**, 14, <https://doi.org/10.1186/2041-1480-5-14> (2014).
21. Golbeck, J. *et al.* The National Cancer Institute's Thesaurus and Ontology. *SSRN Electronic Journal* <https://doi.org/10.2139/ssrn.3199007> (2003).
22. Vasilevsky, N. A. *et al.* Mondo: Unifying diseases for the world, by the world. preprint, Health Informatics. <https://doi.org/10.1101/2022.04.13.22273750> (2022).
23. Ison, J. *et al.* EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics* **29**, 1325–1332, <https://doi.org/10.1093/bioinformatics/btt113> (2013).
24. Gkoutos, G. V., Schofield, P. N. & Hoehndorf, R. The Units Ontology: a tool for integrating units of measurement in science. *Database* **2012**, bas033–bas033, <https://doi.org/10.1093/database/bas033> (2012).
25. De Matos, P. *et al.* ChEBI: a chemistry ontology and database. *Journal of Cheminformatics* **2**, P6, 1758–2946–2–S1–P6, <https://doi.org/10.1186/1758-2946-2-S1-P6> (2010).
26. Brandizi, M. *et al.* Orchestrating differential data access for translational research: a pilot implementation. *BMC Medical Informatics and Decision Making* **17**, 30, <https://doi.org/10.1186/s12911-017-0424-6> (2017).
27. Harris, P. A. *et al.* Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics* **42**, 377–381, <https://doi.org/10.1016/j.jbi.2008.08.010> (2009).
28. Becker, R., Alper, P., Ded, V. & Ebermann, L. ELIXIR Luxembourg Data Information Sheet - DISH. *Zenodo* <https://doi.org/10.5281/ZENODO.7371006> (2021).
29. Patrick Woolley, J. How Data Are Transforming the Landscape of Biomedical Ethics: The Need for ELSI Metadata on Consent. In Mittelstadt, B. D. & Floridi, L. (eds.) *The Ethics of Biomedical Big Data*, vol. 29, 171–197, https://doi.org/10.1007/978-3-319-33525-4_8 (Springer International Publishing, Cham, 2016).
30. Ebermann, L. & Becker, R. ELIXIR Luxembourg Data Hosting Agreement. *Zenodo* <https://doi.org/10.5281/ZENODO.8278663> (2023).
31. Ebermann, L. & Becker, R. ELIXIR Luxembourg Data Use Agreement. *Zenodo* <https://doi.org/10.5281/ZENODO.8279599> (2023).
32. Ebermann, L. & Becker, R. ELIXIR Luxembourg Data Use Responsibilities Agreement. *Zenodo* <https://doi.org/10.5281/ZENODO.8279642> (2023).
33. Vlahou, A. *et al.* Data sharing under the general data protection regulation. *Hypertension* **77**, 1029–1035, <https://doi.org/10.1161/HYPERTENSIONAHA.120.16340> (2021).
34. National institutes of health (NIH). Final NIH policy for data management and sharing and supplemental information. <https://www.federalregister.gov/d/2020-23674> (2020).
35. Joly, Y., Dove, E. S., Knoppers, B. M., Bobrow, M. & Chalmers, D. Data Sharing in the Post-Genomic World: The Experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). *PLoS Computational Biology* **8**, e1002549, <https://doi.org/10.1371/journal.pcbi.1002549> (2012).
36. Abboud, L. *et al.* TEHDAS - WP5 - D5.1 - Report on secondary use of health data through European case studies (2022).
37. Sage Bionetworks. Synapse platform. https://sagebionetworks.org/tools_resources/synapse-platform/ (2024).
38. Foster, E. D. & Deardorff, A. Open Science Framework (OSF). *Journal of the Medical Library Association* **105**, <https://doi.org/10.5195/jmla.2017.88> (2017).
39. Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics* **21**, 615–629, <https://doi.org/10.1038/s41576-020-0257-5> (2020).
40. European Commission. Proposal for a regulation of the European parliament and of the council on the European health data space. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0197> (2022).
41. Carletti, L., Scollen, S., Arenas, J. & Hurst, H. European Genomic Data Infrastructure (GDI): Advancing data-driven biomedical research and personalised medicine solutions to benefit citizens of Europe. *F1000 Research* <https://doi.org/10.7490/F1000RESEARCH.1119428.1> (2023).
42. Jeanson, F. *et al.* Getting Your DUCs in a Row - Standardising the Representation of Digital Use Conditions. *Scientific Data* **11**, 464 <https://doi.org/10.1038/s41597-024-03280-6> (2024).
43. Precisesads326 imi-ju ga 115565 sustainability studies. ELIXIR Luxembourg. <https://datacatalog.elixir-luxembourg.org/e/dataset/ELU-2-E592B0-1> (2017).
44. Courbier, S., Dimond, R. & Bros-Facer, V. Share and protect our health data: an evidence based approach to rare disease patients' perspectives on data sharing and data protection - quantitative survey and recommendations. *Orphanet Journal of Rare Diseases* **14**, 175, <https://doi.org/10.1186/s13023-019-1123-4> (2019).
45. Dyke, S. O. M. *et al.* Consent Codes: Upholding Standard Data Use Conditions. *PLOS Genetics* **12**, e1005772, <https://doi.org/10.1371/journal.pgen.1005772> (2016).
46. Mikkelsen, R. B., Gjerris, M., Waldemar, G. & Sandoe, P. Broad consent for biobanks is best – provided it is also deep. *BMC Medical Ethics* **20**, 71, <https://doi.org/10.1186/s12910-019-0414-6> (2019).

Acknowledgements

This work is partly funded by the contributions of the Luxembourg Ministry of Higher Education and Research towards ELIXIR Luxembourg.

Author contributions

Authors contributions are listed as per Contributor Roles Taxonomy (<https://credit.niso.org/CRediT/>). P.A. Conceptualization, Software, Methodology, Writing – original draft. V.D. Software, Validation, Writing – original draft. V.G., S.H. Conceptualization, Software, Writing – original draft. J.L. Software, Writing – review & editing. D.W. Software, Data curation, Writing – review & editing. L.E. Methodology, Writing – review and editing. S.G., N.B., M.P. Validation, Writing – review & editing. C.T., W.G, V.S. Conceptualization, Supervision, Writing – review and editing. R.B. Conceptualization, Methodology, Supervision. R.S. Supervision.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.A. or V.S.

Reprints and permissions information is available at www.nature.com/reprints.

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



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 Author(s) 2024