Annual Review of Genomics and Human Genetics

Open Data in the Era of the GDPR: Lessons from the Human Cell Atlas

Bartha Maria Knoppers,[1] Alexander Bernier,[1] Sarion Bowers,[2] and Emily Kirby[1]

[1]Centre of Genomics and Policy, School of Biomedical Sciences, Faculty of Medicine and Health Sciences, McGill University, Montreal, Quebec, Canada; email: bartha.knoppers@mcgill.ca, alexander.bernier@mail.mcgill.ca, emily.kirby@mcgill.ca

[2]Wellcome Sanger Institute, Hinxton, United Kingdom; email: sb46@sanger.ac.uk

Annu. Rev. Genom. Hum. Genet. 2023. 24:369–91

First published as a Review in Advance on February 15, 2023

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-101322-113255

Keywords

data sharing, open science, genomics, GDPR, biomedical ethics, data protection

Abstract

The Human Cell Atlas (HCA) is striving to build an open community that is inclusive of all researchers adhering to its principles and as open as possible with respect to data access and use. However, open data sharing can pose certain challenges. For instance, being a global initiative, the HCA must contend with a patchwork of local and regional privacy rules. A notable example is the implementation of the European Union General Data Protection Regulation (GDPR), which caused some concern in the biomedical and genomic data-sharing community. We examine how the HCA's large, international group of researchers is investing tremendous efforts into ensuring appropriate sharing of data. We describe the HCA's objectives and governance, how it defines open data sharing, and ethico-legal challenges encountered early in its development; in particular, we describe the challenges prompted by the GDPR. Finally, we broaden the discussion to address tools and strategies that can be used to address ethical data governance.

1. INTRODUCTION

The twenty-first century has been marked by exponential advances in international data sharing in a number of research fields and industries. The Human Genome Project (33) and subsequent large-scale initiatives, such as the International HapMap Project (21), the 1000 Genomes Project (1), and the Human Pangenome Project (53), required tremendous advances in computational power and collaborations on large, diverse datasets. Building on these projects, the Human Cell Atlas (HCA) is developing a reference map of the cellular compositions of organs and tissues, leveraging recent advances in single-cell gene expression profiling and spatial technologies (34). However, the HCA is distinguishing itself from past initiatives in its more distributed structure and the broad accessibility of its research community (34). As part of its mission, the HCA strives to build an open community that is inclusive of all researchers adhering to its principles and as open as possible (from a regulatory and ethical standpoint) with respect to data access and use.

In this article, we examine how this large, international, collaborative group of researchers is investing tremendous efforts into ensuring appropriate sharing of data as openly as possible. We begin by describing the HCA research community's activities, objectives, and governance; how it defines open data sharing; and some of the ethical and legal challenges encountered early in its development (Section 2). Next, we describe the specific challenges and discussions prompted by the European Union (EU) General Data Protection Regulation (GDPR) (Section 3), illustrating how the HCA approached the issues of compliance and governance while maintaining its open science and data-sharing objectives. Finally, we broaden the discussion on open science to include not only data protection but also ethics governance more generally (Section 4), describing how the HCA Ethics Working Group has provided tools and support to the research community.

2. BUILDING AN OPEN REFERENCE MAP: THE HUMAN CELL ATLAS

2.1. Constructing a Human Cell Atlas

The HCA is an ambitious international data-sharing project. Following in the footsteps of sequencing initiatives such as the Human Genome Project and leveraging recent advances in single-cell genomics, it aims to create a map of every cell type in the human body (44). In doing so, it will provide a baseline for the healthy cellular state, thus enabling the study of disease and illnesses. The HCA has been referred to as Google Maps for the human body, allowing researchers to zoom in and out, from systems down to organs, tissues, cells, and molecules (44, 46). Tissue system networks include experts from the various biological systems represented in the atlas (nervous, peripheral nervous, lymphoreticular, immune, urinary, respiratory, female reproductive, male reproductive, hepato-pancreatic-biliary, gastrointestinal, endocrine, skin, musculoskeletal, cardiovascular, breast, and organoid systems).

The complexity of this endeavor lies in the span of data required to assemble the atlas. The HCA relies on high-resolution measurement on both a cellular level (profiling individual cells and nuclei) and a spatial level (profiling the tissues in context, e.g., within the organ or system) (45). In addition, the types of data used to build the atlas vary in complexity and resolution, including diverse molecular aspects of cells and tissues (their transcriptomes, genomes, epigenomes, proteomes, and metabolomes), structural aspects of cells and their tissues (45), and donor metadata. Furthermore, the atlas requires an understanding of genetic, environmental, and experiential factors contributing to healthy variation across human communities. Unless there is an understanding of the nature and extent of variation among healthy individuals, globally, it will not be possible to detect changes that may lead to diseases being studied (37).

The HCA is ultimately a project requiring both numbers and diversity—of cell types, tissue donors and research participants, local communities, developmental stages (gamete, embryo, fetus,

child, adolescent, and adult), and research groups and networks. Building an atlas that is representative of cell types and states across human populations and regions calls for the involvement of the global research community (36). The success of the project therefore lies in the ability to efficiently and rapidly collect, upload, store, disseminate, and provide access to large datasets.

As part of its open philosophy, the HCA was built as a grassroots initiative, where anyone with an interest in the initiative could sign up to become an HCA member. Nonetheless, because of the complexity of the endeavor, over time, the initiative has developed a more formal structure, alongside its open membership.

The scientific objectives of the HCA are steered and governed by the HCA Organizing Committee. The responsibilities of this committee include convening the community through regular meetings, workshops, and jamborees; coordinating and authoring key documents; defining scientific direction and purpose; providing ethical guidance; defining and upholding processes, including quality-control standards and analytic standards; coordinating the HCA work products; and polling the HCA community at regular intervals for input on issues, including the performance of the Organizing Committee. Several working groups also implement the scientific activities of the HCA; these include the Analysis Working Group, the Standards and Technology Working Group, the Ethics Working Group, and the Equity Working Group. Finally, HCA Biological Networks are communities of collaborators responsible for the study of a particular tissue group, organ, or biomedical theme. Each Biological Network is responsible for efforts related to the imaging, spatial mapping, and single-cell analysis for its focused tissue, organ, or system area of research. In addition, regional networks, including HCA Asia, HCA Latin America, and HCA Africa, ensure that, beyond its working groups, the effort remains connected—and relevant—to scientists internationally. Indeed, to truly reach an open, representative atlas, the HCA community must ensure presence, capacity building, and mutual collaboration across the globe.

Currently, the HCA is structured through two nonprofit legal entities: HCA Inc., established in the United States, and HCA Stichting, incorporated in the Netherlands. HCA Inc. ensures the regulatory compliance of the HCA, as explained in the sections to follow. It also coordinates the administrative and communications activities of the larger HCA project, administers its finances, and governs the HCA Data Coordination Platform (DCP), which includes making all policy decisions concerning the DCP, approving the overall plan for the DCP, and ensuring the plan's successful execution by the major developers of the DCP. In collaboration with relevant scientific, technical, and legal experts, HCA Inc. also oversees the implementation of these policies by providing guidance and making decisions concerning certain key topics, including the definition of the data manifest, the official analysis pipelines, required metadata to reflect data collection standards, the common coordinate framework, and any formal release portal. The activities of the HCA in the EU are structured through HCA Stichting, which is a separate and independent legal entity; these activities include coordinating all HCA EU operations and ensuring that the HCA meets the requirements of the EU data protection compliance mandate.

2.2. Defining Openness Within the Human Cell Atlas

Open science has been an objective sought by many within the research community. It is seen as a potential approach to diminish research bottlenecks, increase data discoverability, reduce access authorization processes, and accelerate research (51). The term, however, can encompass several meanings (3), and the concept is implemented in a number of different ways (open source licensing, open data access, open innovation, etc.) (23).

To maximize the reach and efficacy of sharing datasets, the HCA adopted an open science model early on. The open philosophy of the HCA encompasses several facets of its organization. All interested researchers, from any career stage or geographical region, who share the HCA values

are welcome to join the HCA community and participate in its activities through its publicly available communication channels and collaboration platforms (44).

In terms of data access, the HCA fosters access that is as open as possible while being cognizant of ethical or legal requirements that place certain limitations on data sharing. Striking a balance between sharing data widely through open access, as opposed to controlling access to certain types of data of a more sensitive nature, is crucial to the HCA's objective of building a widely accessible map of the human body and its cells. Indeed, certain types of biomedical data and metadata are considered more sensitive than others, due to the type of information they reveal (e.g., whole genomes, medical record information, or potential stigmatizing personal or populational characteristics). In addition, privacy risks may be increased due to the potential of reidentification through the linkage of datasets existing in different databases (26, 29)—a risk that may be exacerbated if certain data are freely or publicly accessible.

2.3. Ethical and Legal Challenges

Open access can pose certain challenges with respect to ethico-legal issues. For instance, being a global initiative, the HCA must contend with a patchwork of local and regional privacy rules. A notable example of this complexity emerged with the entry into force of the EU General Data Protection Regulation (GDPR) (19) in 2018, which caused a fair amount of concern in the biomedical and genomic data-sharing community (47). Indeed, a report titled "International Sharing of Personal Health Data for Research" published by All European Academies (ALLEA), the European Academies' Science Advisory Council (EASAC), and the Federation of European Academies of Medicine (FEAM) (2) highlighted GDPR-caused barriers and disrupted data streams. The report went so far as to state that the GDPR was harming EU leadership. This call for GDPR regulatory reform to facilitate reciprocity in sharing health data is not new (8), but perhaps sharing some of the mechanisms and tools to mitigate the effect of the GDPR is warranted. Section 3 further examines the HCA's experience with the application of the GDPR to its research activities as well as key governance tools that it adopted.

In addition to data protection concerns, traditional ethical, legal, and social issues (consent, sampling of human tissues, etc.) have also been a challenging aspect of the HCA's design. Indeed, due to its very nature, the project requires a wide breadth of tissue-sampling strategies. For instance, from a developmental perspective, the HCA relies on the collection of tissue from gametes, embryos, fetuses, children, adolescents, and adults (28, 49). Because of different degrees of vulnerability or social or cultural sensitivities, the sample procurement and consent requirements vary extensively among these different sampling groups. Furthermore, because not all cells in the human body can be sampled from living, healthy individuals, different tissue procurement scenarios need to be envisaged for both living and deceased donors—including use of clinical leftover tissue samples, postmortem organ or whole-body donations, and so on. Again, this means that a number of different regulations, policies, or ethical standards will apply. Needless to say, in all of the examples above, regulations and requirements may vary from country to country, making a global initiative such as the HCA quite complex from a regulatory and ethical standpoint. Moreover, matters of both open data sharing and data protection require consideration in these different tissue-sampling scenarios, further complexifying the landscape. In Section 4, we highlight some of the work that the HCA Ethics Working Group has undertaken to provide HCA researchers with resources and tools to implement in their local tissue-sampling initiatives.

3. THE DATA PROTECTION PARADIGM

The GDPR is a sweeping EU privacy law that entered into force in 2018, designed with the intention of creating a harmonized EU and European Economic Area (EEA) framework to regulate the

use of personal data (19), although the resulting legislation is not completely harmonized. This legislation uses a combination of guiding principles, procedural rules, and flexible enforcement powers to create a holistic regulatory regime. It has attracted widespread international attention due to its potential to impose significant administrative fines on regulated parties (19). The GDPR has proved to be a popular model for legislatures worldwide, and similar legislation is in the process of being enacted in numerous non-EU/EEA countries (24, 25).

The GDPR builds on prior legislative and regulatory efforts both within and outside of the EU and EEA. One notable example is the Organisation for Economic Co-operation and Development (OECD) privacy guidelines (38). First drafted in 1980 and updated in 2013, these guidelines established the foundational principles that undergird contemporary privacy and data protection norms recognized around the world. In 1995, the European Commission implemented the Data Protection Directive (DPD) (18), an EU-wide framework requiring EU Member States to integrate certain foundational guarantees of data protection and privacy into their domestic legislation. The major elements of the DPD are replicated in near-identical form in the GDPR, providing a measure of guidance regarding GDPR compliance best practices, despite the recent adoption of the latter law. Though the GDPR remains mostly untested before the courts, the jurisprudence and regulatory guidance applicable to the DPD can help in the interpretation of the GDPR. In 2017, the OECD adopted the Recommendation on Health Data Governance (39), which encourages nations to adopt legislation enabling the processing and sharing of health data.

Other commentators have authored an extensive literature identifying practical difficulties for research consortia in ensuring data protection compliance, including critiques directed at the GDPR. Our ambition is instead to provide practical guidance to assist biomedical research data repositories in navigating such challenges, building on the experience of the HCA. To this end, we describe certain difficulties that research consortia face in achieving GDPR compliance and provide actionable guidance to help them navigate these issues. This guidance builds on the experience of the HCA.

The GDPR applies to controllers and processors that process identifiable personal data. Its ambit is usually limited to the processing of data by entities that are established in the EU or the EEA. In some limited circumstances, the GDPR can find application outside of the EU/EEA—for example, if entities outside the EU/EEA target their services to persons in the EU/EEA. The language of the GDPR uses the concepts of controller and processor to describe the persons and legal entities that it regulates in their use of data, ascribing different responsibilities to each such role (19). The language of the GDPR uses identifiable personal data as the litmus test for determining which data are regulated and which are not (19). Therefore, in determining whether the GDPR applies to a particular actor, two consecutive analyses must be performed: first, determining whether the data being processed constitute identifiable personal data within the context of their use, and second, determining whether the specified actor is behaving as a controller, a processor, or neither, relative to the concerned identifiable personal data. Strategies for addressing these considerations are discussed in the following sections. The final sections go on to address other issues relating to GDPR compliance.

3.1. Joint Controllership and the Structure of Research Consortia

A first point of tension between the structure of biomedical research consortia (such as the HCA) and the assumptions implicit in data protection legislation lies in the GDPR concepts of controllers, joint controllers, and processors. The GDPR uses these three categories to establish the respective roles and responsibilities of different legal entities and natural persons engaged in the processing of personal data.

Controllers are the entities that determine "the purposes and means of the processing of personal data" (19, article 4). This is in contrast to processors, which perform "any operation or set of operations. . .on personal data" (19, article 4). In simple language, the controllers determine how data will be used, and the processors provide the technical tools or associated services required to use the data as intended. Joint controllers are controllers that collaborate in determining how data will be used.

From the standpoint of liability (i.e., legal responsibility for upholding obligations), the GDPR adopts the following approach. Controllers are responsible for upholding the principles and the procedural responsibilities established in the GDPR and can be subject to administrative fines or civil liability for failure to do so. Joint controllers are held jointly and severally liable for breaches of the GDPR that other joint controllers collaborating with them cause. Jointly and severally liable is a legal term that means that each of the joint controllers can be held responsible for the total penalty owed regardless of their respective fault or contribution to the damages caused, which is held to be the case unless a particular joint controller can demonstrate that it is "not in any way responsible for the event giving rise to the damage" (19, article 82). This liability means that in a large, complex research collaboration among multiple institutions (as is often the case for large-scale biomedical research consortia), legal consequences for the large-scale harms that one participant causes could be imposed in full on any one of the other participants (19, 32).

Processors, for their part, are held liable for breaches of the more limited GDPR obligations that are directed to processors and for acts that contravene the instructions of the controller or controllers. The GDPR therefore places the balance of the legal responsibilities for ensuring the appropriate use of data on the controller or controllers that direct how data will be used, rather than on the service providers or partners that simply implement their instructions.

As stated above, biomedical research consortia are often structured as large, informal networks of academic, clinical, industry, and individual collaborators. This structure is adopted for numerous reasons. The first is that such research efforts draw from highly specialized and diverse pools of collaborators that are located in numerous different jurisdictions and institutions. The composition of such networks is not stable but rather differs from one consortium to another, with distinct arrangements being better suited to some research proposals than others. Therefore, relevant individual or institutional stakeholders benefit from remaining within their respective institutions and collaborating ad hoc on a project-specific basis.

It is often the custom—and preference—of biomedical research consortia to define the roles and responsibilities of each of the specified consortium collaborators in clear language, as part of multiparty contracts. This approach ensures that each participating individual or institution is held responsible for the acts, services, and technologies belonging to its defined role but not those of other collaborators, which has two purposes. First, it encourages potential collaborators that offer specialized services but do not receive remuneration or possess the resources to bear risk to participate in the consortium, conditional on limits being placed on their prospective legal liability. Second, it encourages collaborators to invest resources in mitigating the risks associated with their own contributions to the overall data-processing network. As each collaborator bears the agreed risks associated with their respective contributions to the overall consortium, collaborators have ample incentive to ensure that their allocated risks do not materialize.

There is therefore a contrast between the GDPR approach to the assignment of responsibilities between controllers and the approaches that research consortia have historically favored. Research consortia prefer to delimit the legal responsibilities of distinct collaborators using contracts that establish clear boundaries to the respective responsibilities of each, which encourages the participation of more collaborators, enhancing the potential for improved data protection compliance

and heightened data utilization because of the presence of a larger number of expert contributors. Parties will be incentivized to collaborate in the consortium and to be effective in their functions because their legal responsibilities mirror the responsibilities the consortium has entrusted to them. Conversely, the GDPR enhances the potential for enforcement, accountability, and compensation for both rights-bearers [referred to as "data subjects" in the GDPR (19)] and regulators. Because collaborators in a research consortium can be held legally responsible for acts outside of their formal responsibilities in the consortium, the GDPR has the potential to dissuade participation in consortium development due to the real and perceived risks of legal liability arising from such participation (19, 32).

It is possible that, in the future, special-purpose legislation will be implemented to create less onerous liability rules in the healthcare, public-health, and biomedical research sectors. Until then, research consortia that create long-term infrastructure that is intended to host data shared by upstream data contributors and downstream data users should take care to enact the following three measures.

First, it is recommended to define the respective obligations of each of the consortium stakeholders using contracts. Controllers and/or joint controllers should be identified as such and their respective responsibilities made clear. Processors should likewise be identified and their obligations established using contracts. Entities that are not intended to act as either controllers or processors should also be defined as such using contracts. These obligations can also be established in a data protection impact assessment or other central document dedicated to data protection compliance. For partners that are external to the consortium itself, these roles and responsibilities can instead be confirmed in the contracts that establish their respective roles and duties.

In addition, collaborators should ensure that their actions in the consortium are aligned to their purported GDPR roles. Entities that do not act as controllers should not bear formal authority to determine the "purposes and means" of data processing, and entities that do not act as processors should not process data. Collaborators that are neither controllers nor processors and instead act in advisory or supporting capacities should ensure that their roles are structured in consultative or delegated capacities rather than as autonomous decision-making bodies. To this end, research consortia that engage parties to perform specified functions should ensure that the responsibilities entrusted to them are aligned with their intended GDPR roles and should stipulate the anticipated GDPR roles of such parties in the concerned contracts (19, 32).

The second recommended measure is to centralize data controllership responsibilities in a designated EU or EEA legal entity. For some research consortia, this entity could be an institution that is designated as the data steward among the collaborating institutions; for others, it could be an EU or EEA legal entity that is created to represent the consortium and acts as the controller for GDPR purposes. This legal entity should perform all of the functions that the GDPR ascribes to controllers, to the exclusion of all other entities, lest these other entities be characterized as joint controllers. Clear contracts should be stipulated with upstream data depositors contributing data to the consortium, establishing the conditions of such controller-to-controller data transfers. The same should be done with downstream data users, establishing that their obtaining such data constitutes a similar controller-to-controller data transfer. The designated controller should, as stipulated, enter into the necessary contracts with data processors that act on its behalf.

Finally, consortia that include both EU/EEA and non-EU/EEA participating institutions should appoint an EU/EEA institution to be responsible for data controllership and data stewardship in the EU/EEA as well as an independent non-EU/EEA institution that holds data stewardship responsibilities for the consortium's activities outside the EU/EEA. Section 3.4 discusses additional strategies for including non-EU/EEA collaborators.

3.2. Assessing the Legal Identifiability of Biomedical Research Data

A second significant challenge that research consortia must face in structuring their compliance with the GDPR is to distinguish identifiable personal data, which the GDPR regulates, from nonidentifiable (i.e., anonymized) or nonpersonal data, which it does not regulate. The GDPR imposes certain requirements on the processing of identifiable personal data in general and imposes more stringent, additive requirements on the processing of identifiable personal data that fall under special categories enumerated in the GDPR. In addition to not applying to nonidentifiable or nonpersonal data, the GDPR does not apply to the data of deceased persons, though EU and EEA Member States can implement domestic legislation to extend its application to such data (19).

Information is considered to be identifiable personal data if the controller, or a proximate third person, can use a method that is "reasonably likely to be used" to identify the person to whom those data relate (19, recital 26). Methods of reidentification that are theoretical but could become viable in the future, or that are impracticable to implement due to practical challenges or cost-intensiveness, are not considered. In this respect, the GDPR uses a contextual assessment to determine whether data should be considered identifiable personal data. This assessment considers both the nature of the data (i.e., the presence or absence of potentially identifying features) and the risk of reidentification that arises in the circumstances of its use (6, 9, 19, 41).

This assessment is important for the design of data access controls (e.g., controlled access, registered access, and open access portals) that enable prospective users to access or request access to consortium datasets (12). Efforts to develop platforms for infrastructure science, such as the HCA, are required to grapple with complicated legal determinations such as assessing the identifiability of the data that such platforms ingest from data contributors. The assessment of data identifiability must be performed for all categories of data that such platforms ingest and share, each of which presents unique challenges. Performing these assessments is cost-intensive, both because continuing debate exists as to the breadth of the legal test to be performed and because significant scientific expertise is required to assess the reidentification risk associated with different categories of biomedical data.

Let us now consider three distinct examples of potential identifiable personal data typically encountered in biomedical research. The first is record-level data and/or metadata that refer to a single individual who contributed data to a local research project, which then deposited its research data in a biomedical data repository; this type of data can include structured data (such as data extracted from clinical records) or demographic and sample-specific metadata about a tissue sample and the relevant donor. The second is unstructured or qualitative data derived from a single individual, which would include genetic data derived from an individual's whole-genome sequence (a small portion, rather than the full whole-genome sequence or a significant portion thereof). The third is aggregate information derived from multiple research participants, such as allele count frequencies or reference maps that aggregate the data of multiple research participants to create visual representations thereof or statistical data at the group level. In each of these cases, the analysis of data identifiability accounts for both the characteristics of the data (i.e., the presence or absence of direct or indirect identifiers) and the reidentification risks arising from the context of its use.

3.2.1. Record-level data and/or metadata. Organizations can take several measures to assess and reduce the identifiability of structured record-level data about individuals. A combination of quantitative and qualitative methodologies can assess the potential for such data to be used to perform the reidentification of an individual, alone or in combination with other available data (13). The objective of this assessment is to obtain a relatively accurate, albeit imperfect, measure of the reidentification risk associated with the data.

Quantitative tests of data identifiability include *k*-anonymization. This method divides the variables in a dataset into direct identifiers, indirect identifiers, and presumed nonidentifiers. Direct identifiers, such as name, civic address, and social insurance number, are sufficient to identify an individual on their own. Indirect identifiers can, in combination, provide sufficient contextual information about an individual to enable their reidentification; examples of indirect identifiers include age, gender, or ethnicity. Presumed nonidentifiers generally do not enable an individual to be reidentified, because they cannot be compared to or combined with other external data elements to enable individual reidentification; examples of these elements include select biomarkers or clinical measurements that change across time and are therefore unlikely to be replicated across multiple databases (6, 13).

To estimate the reidentification risk associated with a dataset and reduce it through deidentification methods, the following methodologies should be used. First, direct identifiers should be removed from the dataset and/or replaced with a code; if direct identifiers remain in the dataset, the data should be considered identifiable personal data. Second, the presumed nonidentifiers, which do not contribute to the estimated reidentification risk, should be labeled as such. Third, the indirect identifiers in the dataset should also be labeled as such. Then, for each record in the dataset, the number of other records that have the same combination of indirect identifier values should be counted. If each record's combination of indirect identifier values is shared with at least 10 other records—meaning that there are never fewer than 11 records that are indistinguishable from the standpoint of data identifiability—then the dataset can often be considered anonymized and suitable for public release (6, 13).

The number 11 was chosen somewhat arbitrarily but has received support in the guidance of numerous regulators and technical experts worldwide. GDPR regulators have not yet taken a position for or against the appropriateness of this quantitative threshold, but it is beginning to emerge as an international consensus position (6, 13, 17). There are numerous ways to reduce the reidentification risk of data below this threshold, including the redaction of certain fields, the generalization of variables, and the addition of noise to certain variables (13, 17).

For controlled data releases, it is sometimes possible to consider data anonymized even if some records share an identical combination of indirect identifier values with fewer than 10 other records. There is an extensive literature on the quantitative and qualitative metrics that enable the assessment of data identifiability according to different models of controlled data release. These methods first evaluate the reidentification risk associated with the data in pure quantitative terms and then amend it to account for mitigating or aggravating factors arising from the data's release model, the safeguards implemented, and the incentives for third parties to attempt reidentification. A full discussion of this literature is outside of the scope of this article (13, 17).

Generally, the methodology chosen to estimate residual reidentification risk necessarily relies on several subjective and qualitative assumptions. First, it is necessary to distinguish potential indirect identifiers from presumed nonidentifiers, and although some such determinations are evident, others could prove contentious or erroneous. Second, the choice of the threshold of records with analogous identifiers that is required for the data to be considered nonidentifiable is based on a subjective assessment. Third, the mitigating role ascribed to the incentives and disincentives to perform reidentification, to the access controls implemented, and to the technical audit or organizational metrics utilized to reduce the risk of reidentification in practice is also based on contextual assessments of a subjective nature (6, 13).

For unstructured data or data that are highly variable from one individual to another, the following heuristics can prove useful. In assessing the identifiability of such data, data protection experts and scientific experts should collaborate in determining whether the circumstances of the data's release and the data's technical characteristics create a foreseeable prospect of

reidentification (48). Technical measures can be implemented to reduce the residual risk of data reidentification. These measures should be selected so as to best preserve the scientific utility of data. Clear records should be maintained of the deidentification methods considered and of their potential negative effects on the quality, utility, and fidelity of the data. The method of deidentification that achieves the best balance between maintaining data utility and reducing the risk of deidentification should often be preferred (6).

3.2.2. Individual-level unstructured data. The second example is individual-level unstructured data. It is true that, for example, individual-level genetic data, imaging data, and neuroimaging data often contain elements that are unique to a single individual or a small number of persons from among the general population. Nonetheless, only certain features of these data could enable the reidentification of individuals, while others likely could not (22). Image headers and certain facial features in brain scans could potentially enable individual reidentification, and the removal of these elements eliminates data protection risks while maintaining data utility, creating data that can most likely be considered anonymized (5, 22).

For genomic data, conversely, limited information about an individual's genetic variants could have a low risk of individual reidentification, because there is no available comparator dataset that someone could use to carry out reidentification, especially if such data are subject to controlled release to trusted parties. Furthermore, modifying such data could reduce its scientific utility (6). If genomic data are modified to reduce reidentification risk, then it may be prudent to retain a greater proportion of the data's individual-specific features, especially those that could not realistically be cross-referenced against external comparator data for the purpose of performing reidentification. Assumptions regarding the kinds of external comparator datasets that are available to attempt reidentification might need to be revised in the future, as databases of public genomic information and other specialized databases become more common. The amount of effort required to attempt reidentification, in terms of specialized equipment and labor, should also be considered (9).

3.2.3. Data aggregated from multiple individuals. The last example is data aggregated from the records of multiple individuals. Though aggregate data are not often considered to be identifiable personal data, it can be worthwhile to consider the circumstances in which aggregate data can reveal individual-level information. Methodologies such as differential privacy, and derivatives thereof, assess how much information unique to a specific individual contributed to an output result, such as a statistical analysis (11, 42). The reidentification risk considered is that multiple such analysis results could be compared with one another to infer the presence or absence of a particular individual in a dataset. To safeguard against such reidentification risks, features such as query budgets or noise addition can be integrated into data analysis platforms. In most instances, such precautions are not required, because the costs of attempting reidentification from aggregate data are high and the incentives to do so are limited or nil (43).

In determining an approach to data deidentification and data release, data custodians should first assess the prospect for reidentification to occur. This assessment should leverage empirical knowledge such as quantitative tests and/or specialized information relating to the scientific nature of the concerned data. The identifiability analysis should be followed by the implementation of access controls that are appropriate to the determination made. Identifiable data should often be held in controlled access unless necessary legal and ethical authorizations are obtained to release it in open access. Anonymized data can often be released in full open access, without implementing data access controls (12).

Last, data deidentification methods can be used to reduce the identifiability of data, which might succeed in rendering otherwise identifiable data anonymized. The trade-off between

rendering data anonymized (and therefore enabling a more open release model) and maintaining the data's scientific utility (possibly requiring a controlled release of data and entailing a heightened GDPR compliance burden) should be considered.

3.3. Legal Justifications for the Processing and International Transfer of Data

The third major challenge to be addressed in processing personal data in the ambit of a biomedical research consortium is to identify appropriate legal justifications for both data processing and international data transfer.

Much of the data that biomedical data repositories ingest could be considered identifiable personal data, as well as special-category data within the larger umbrella. To process identifiable personal data, data controllers must fulfill one of the conditions (referred to as lawful bases) established in article 6 of the GDPR. In addition, to process identifiable personal data that are considered part of a special category, data controllers must fulfill one of the conditions established in article 9 for those types of data (19).

Certain general rules are applicable to the selection of a lawful basis. A controller often cannot change lawful bases for a continuous act of data processing if the lawful basis that it had previously relied on ceases to be applicable. However, if a controller shares data with a downstream controller to use for its own purposes, then it is normal and expected that the two controllers could rely on different lawful bases to legitimate their respective data-processing activities.

The use of some lawful bases must be supported by an enabling provision established in the domestic legislation of the concerned EU/EEA Member State, rather than through simple reliance on the text in the GDPR—that is, Member State law must explicitly or implicitly empower the concerned controller to use the data for such purposes. Some lawful bases are directed to specified categories of controller alone (e.g., some are directed to public-sector entities, while others are directed to entities processing data in a private-sector context) (19, 32).

Biomedical research consortia often operate across numerous EU/EEA Member States and often include non-EU/EEA partners. Furthermore, their activities may include the participation of public-sector bodies, private-sector bodies, and numerous other actors in the health sector, including public-health agencies, laboratories, and clinicians. The implications of this structure for the identification of appropriate lawful bases according to the GDPR and EU/EEA Member State law are discussed here; the implications for non-EU/EEA partners are discussed in Section 3.4.

The broad participation that is common in research consortia can create challenges in identifying an appropriate lawful basis to enable the processing of identifiable personal data and/or special-category data. These difficulties arise because the lawful bases are structured to meet the needs of data controllers that are established in a specific EU/EEA Member State or that have a specified role in the public sector, private sector, or health sector, but biomedical research data repositories often straddle boundaries with regard to both their jurisdictional locus and the nature and role of their constituent entities.

One potential solution to this conundrum is for biomedical research data repositories to incorporate in the EU or EEA, which creates a relatively clear jurisdictional tie to a specified Member State. The corporate form that such repositories adopt can also help determine which lawful basis will be the most appropriate for their intended data-processing activities. In acting as a singular legal entity rather than as an amorphous group of collaborators that may be characterized as joint controllers, this legal entity can establish clear boundaries to its controllership. Contracts can be used to define the circumstances in which external data controllers contribute data to the repository and the circumstances in which the repository provides data to other controllers. The beginning and end of each controller's activities thus become unambiguous (19).

These strategies can help ensure that biomedical data repositories are able to make an informed selection of a lawful basis to inform their data protection compliance activities. It nonetheless remains an open question which lawful bases are appropriate to justify the safeguarding of data in biomedical research data repositories.

The selection of an appropriate lawful basis is not a simple legal formalism that acts as a necessary precondition to data processing. Some lawful bases restrict the purposes for which data can be processed, require performing additional rights-balancing tests prior to data processing, or require the ongoing actualization of data-subject rights in all instances. To summarize, the selection of an appropriate lawful basis can determine the procedural and substantive conditions that must be respected in processing the data (19).

Article 6 of the GDPR establishes a large number of distinct lawful bases to process personal data in general. Three principal lawful bases that might prove useful to large biomedical research data repositories are the following (19):

1. Legitimate interest: This lawful basis could be suitable for controllers that are not public authorities or for public authorities acting in a capacity other than the performance of their public tasks. The GDPR specifies that this lawful basis is not available to public authorities that are processing data in the performance of their tasks. It also requires those controllers to determine that "the interests or fundamental rights and freedoms of the data subject" do not "override" the legitimate interest that the controller pursues.
2. Public interest: This lawful basis enables controllers to process personal data where doing so is "necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller." The conditions according to which such processing can be performed, however, must be further specified in EU law or Member State legislation. This can create challenges for controllers that wish to rely on this lawful basis prior to its actualization in Member State law and can hamper its uniform use for data-processing activities that are performed across multiple Member States.
3. Consent: Controllers can process personal data on the basis of a consent that the data subject provides. Because consent holds a sacrosanct place in both biomedical research and clinical care, it at first appears to be a natural fit for data processing in biomedical research repositories. However, there are several challenges to this position. The first is that GDPR consent must be specific in nature, whereas biomedical research increasingly recognizes reliance on broad consent to the future research use of data. Furthermore, an imbalance of power is often considered to occur between public authorities and data subjects. The European Data Protection Board, an institution composed of EU supervisory authorities and tasked with issuing interpretive guidance, has also issued publications on this topic in which it confirmed that informed consent obtained for research purposes in the context of a clinical trial might not always be sufficient to meet the requirements of GDPR consent (15). For these reasons, it can be challenging for biomedical data repositories to use consent as the lawful basis that underpins their data-processing activities.

Article 9, which applies to the processing of special categories of data, also offers several relevant options (19):

1. Consent: The challenges inherent in invoking consent as a lawful basis for the processing of personal data under article 6 of the GDPR also arise in attempting to invoke it as a lawful basis for the processing of special-category data under article 9. In addition, consent relied on to process special-category data must be explicit consent to the intended acts of data processing. This could further inhibit reliance on the broad consent that is often used to

obtain the consent of research participants to the use of their data for unspecified future research purposes.

2. Data manifestly made public by the data subject: This lawful basis appears promising for citizen science efforts and other efforts to host user-generated or user-submitted content on public repositories. Its boundaries have not been explored in the literature or in the guidance of regulators. One potential challenge for its use in hosting data generated in research contexts or clinical settings is that the data contributor is often a member of research, clinical, or health institution personnel acting on behalf of the data subject. It is not clear whether the subject's agreement to deposit data on a public platform, in the context of research or clinical care, constitutes making such data public.
3. Scientific or historical research purposes or statistical purposes in the public interest: This lawful basis, on first reading, appears ideal for biomedical data repositories; however, there are a few practical challenges that detract from its initial appeal. First, the lawful basis must be actualized in enabling Member State legislation before controllers can rely on it. Second, the functioning of the lawful basis will differ across the laws of different Member States, which can create challenges for international biomedical data repositories that operate across numerous countries and hope to place uniform reliance on this lawful basis. Similarly, the GDPR provision that enables the processing of special categories of data for scientific research purposes must also be implemented at the level of the Member State, and because most large-scale data-driven biomedical research is international in nature, it could be unclear which Member State law, or combination of Member State laws, would be appropriate.

This analysis of the lawful bases available in the GDPR demonstrates that none appears to align with the needs of biomedical data repositories in an elegant manner. Nonetheless, a number of the available lawful bases will likely prove sufficient for biomedical data repositories. Incorporating in a specific Member State can help to guide the selection of enabling Member State legislation where required, and adopting a clear corporate form can also help to determine which of the public-sector-oriented and private-sector-oriented lawful bases might be most appropriate for a particular biomedical data repository.

3.4. International Transfers of Personal Data

The GDPR has a special legislative regime that applies to the outbound transfer of data from the EU/EEA to non-EU/EEA jurisdictions. This regime of international data transfers requires controllers that transfer data to other controllers outside the EU/EEA to demonstrate that a justification exists for such transfers, and in certain circumstances, additional safeguards must be implemented as a precondition to performing the transfers. The logic underpinning these requirements is that both the GDPR and the fundamental constitutional rights of EU citizens entitle individuals in the EU/EEA to benefit from GDPR protections, and the GDPR international transfer mechanisms are therefore intended to ensure that international transfers of personal data from the EU/EEA do not deprive data subjects in the EU/EEA of those guarantees (27). This governing logic helps to understand the following structure of international data transfer rules, which on a plain reading appears arcane.

In the absence of a justification, the GDPR presumes that transfers of data from inside the EU/EEA to outside the EU/EEA are unlawful; to perform such a data transfer, a controller is therefore required to have a legal justification for doing so. The default legal justification is that the EU/EEA has determined that the recipient jurisdiction provides an adequate standard of data protection, meaning that the European Commission has performed a review of the data protection

legislation, privacy law, surveillance and policing practices, and effective access to legal recourse that are available in the jurisdiction of destination. After performing this review, the European Commission can conclude that the combination of foregoing elements provides an adequate standard of data protection at the destination of the concerned data transfer. It then issues an adequacy decision that describes the breadth of the jurisdictions, geographical area, international organizations, and/or economic sectors to which it applies. It also sets out any additional conditions applicable to the data transfer and to the use of data after the transfer that the European Commission considers necessary to uphold the standard of data protection guaranteed by EU/EEA law (19).

At present, adequacy decisions that benefit 14 countries or subnational territories have been issued. This enumeration does not include the now-defunct adequacy decisions that had previously benefited commercial organizations in the United States that self-certified as compliant (27), because the Court of Justice of the European Union invalidated these adequacy decisions. Because of the small number of jurisdictions that benefit from favorable adequacy decisions, it can be assumed that the majority of international data transfers from the EU/EEA will not be performed on the basis of an adequacy decision. The GDPR also establishes numerous mechanisms for international data transfers other than adequacy decisions, including binding agreements between public bodies, binding corporate rules, standard contractual clauses (SCCs), codes of conduct, and certification. Of these transfer mechanisms, the majority are not available at present because the European Commission must review and approve the tools prior to their use (27).

SCCs are the sole international data transfer mechanism available at present. These clauses are to be integrated into the contracts governing data transfers between an EU/EEA controller and a non-EU/EEA data recipient. They impose the principal requirements of the GDPR on the data recipient and translate them into enforceable contractual commitments, providing a cost-effective mechanism for performing personal data transfers from the EU/EEA to third countries.

In the context of biomedical data repositories, there are two major caveats to this position (19). The first is that numerous academic and research institutions in the United States (a major recipient of outbound data transfers from the EU/EEA) cannot sign the SCCs, because US domestic law prohibits these entities from binding themselves to contractual clauses that engage their liability, as the approved SCCs do (35). Other third countries and international organizations might also be subject to similar restrictions. The second caveat is that entities that transfer data from the EU/EEA to third jurisdictions that do not benefit from an adequacy decision are required to perform an assessment of the domestic law and local governmental practices in the destination jurisdiction prior to performing such transfers. This assessment verifies whether the law, the practices of authorities and surveillance bodies, and the courts of the destination jurisdiction are capable of ensuring that the fundamental rights of EU citizens are upheld. If they are, then the transfer can proceed unfettered; if they are not, then it is necessary to implement supplementary measures to raise the standard of data protection in the recipient jurisdiction to meet the standard established in EU data protection and fundamental rights legislation (10). These measures can be of a legal, organizational, or technological nature. Often, technological measures precluding the surreptitious use of information will prove to be necessary in circumstances where local surveillance bodies and police authorities breach the rights of EU citizens and/or where judicial recourse for the breach of fundamental rights and data protection rights is not available (16).

In sum, therefore, the most cost-effective path to data protection compliance in international data transfers for biomedical data repositories is to rely on standard SCCs to perform these outbound data transfers. This will require an analysis confirming that law and practice in the destination jurisdiction are capable of ensuring respect for the fundamental rights of EU citizens.

If the analysis reveals that the law and practice cannot meet such standards, additional measures must be implemented to raise the standard of data protection. Entities that cannot sign the SCCs also cannot rely on this transfer mechanism.

Biomedical research repositories that are incorporated in the EU/EEA can consider also incorporating an independent non-EU/EEA entity to conduct their non-EU/EEA data-processing activities and other non-EU/EEA functions. Using this approach, EU/EEA entities can transfer personal data to their non-EU/EEA counterpart with SCCs, and the non-EU/EEA entity can then perform downstream transfers to third parties that cannot bind themselves to the SCCs. This is the case because the recipients of international data transfers made using SCCs do not need to use SCCs to further transfer data to third parties (14).

While the GDPR poses a number of challenges for biomedical research consortia, several strategic governance approaches can be adopted to adhere to these requirements while fostering good data management practices. These challenges and proposed solutions are summarized in **Table 1**. In the next section, we propose additional governance approaches and tools that the HCA has adopted to foster adherence to legal and ethical best practices.

4. GOVERNING (OPEN) DATA

4.1. Building Proportionate Governance

In addition to adopting the foregoing strategies with respect to GDPR regulatory compliance, from a general ethical and legal standpoint, developing adequate data governance frameworks and tools ultimately serves to facilitate data contribution and access. Indeed, tools can be developed to foster the contribution of datasets that fulfill minimal technical requirements set by the consortium and can be shared and used for a common purpose (7).

However, in the context of open access to biomedical data (particularly omics data), efforts must be made to carefully review datasets (or categories of datasets) made available by the consortium and to justify their level of openness. While data protection regulations do not preclude open, unrestricted sharing of data, such regulations do often require data custodians to undertake a thorough assessment of data types and determine what data can be shared, with whom, and how. Furthermore, beyond the data protection regulatory issues described in Section 3, there are ethical governance points to consider in the implementation of open science, particularly open data sharing. While at first it may seem counterintuitive that a governance framework is required to provide free, unfettered access to data, experience with the HCA has shown that this is a useful way to strengthen open science principles.

4.2. Helping the Scientific Community: The Human Cell Atlas Ethics Toolkit

Formally established in September 2018, the HCA Ethics Working Group is tasked with discussing the ethico-legal issues relevant to the HCA, particularly in relation to its global reach, and providing input on ethics governance documents and tools for the HCA. The working group is currently composed of 15 members and 15 observers (including funder representatives, members of other working groups, etc.), spanning 12 countries. The Centre of Genomics and Policy at McGill University was funded to coordinate the activities of the Ethics Working Group, including leading the development of a suite of tools to help HCA researchers with implementing the HCA within their own local ethics policy frameworks. The majority of tools in this Ethics Toolkit (available at **https://www.humancellatlas.org/ethics**) were initially developed to prioritize open data sharing, but have also been adapted to situations where controlled (managed) access may be required, and appropriate language will be included across the different documents.

Table 1 Regulatory compliance challenges and proposed solutions

Regulatory compliance challenge	Proposed solution
Distinguishing GDPR joint controllers, controllers, and processors in large-scale biomedical research consortia	The following measures are recommended: ■ Use bilateral contracts between parties and a master contract or master document to establish the responsibilities of each collaborator. ■ State the intended GDPR role of each collaborator in the relevant contract or contracts. ■ Ensure that each collaborator respects the legal definition of their intended GDPR role through their actions. It may also be advisable to create a dedicated EU/EEA legal entity that is responsible for acting as the designated data controller for GDPR-regulated data.
Performing international transfers of personal data from the EU or EEA to third countries	Personal data can be transferred to international data transfer recipients that benefit from an adequacy decision without additional precautions being implemented. For international data transfers directed to jurisdictions that do not benefit from an adequacy decision, the following steps must be taken: ■ Identify a GDPR transfer mechanism that enables the transfer of the data outside of the EU/EEA. The only mechanisms presently available are the SCCs that the European Commission have approved. ■ Perform an assessment of the domestic law and local governmental practices in the destination jurisdiction prior to initiating such transfers. This assessment verifies whether law and practice in the destination jurisdiction are capable of ensuring respect for the fundamental rights of EU citizens as recognized in EU constitutional law. ■ If it is determined that the law and practice in the destination jurisdiction are not capable of ensuring respect for the fundamental rights of EU citizens, implement supplementary measures when performing the data transfer to better safeguard such rights. These measures can be of a legal, organizational, or technological nature. Often, technological measures precluding the surreptitious use of information will prove necessary in circumstances where local surveillance bodies and police authorities breach the rights of EU citizens and/or where judicial recourse for the breach of fundamental rights and data protection rights is not available. This is the case because legal (e.g., contractual) and organizational measures are not likely to succeed in precluding the state from compelling access to personal data and thus breaching the fundamental rights of EU citizens.
Performing international transfers of personal data from the EU or EEA to recipients who cannot sign the SCCs	Certain categories of organizations cannot sign the boilerplate SCCs that the European Commission has approved. For example, many research institutions in the United States cannot sign such clauses because statutes preclude them from agreeing to select terms in the SCCs, such as the clauses on liability. Because the recipients of international data transfers made using SCCs do not need to use SCCs to further transfer data to third parties, one approach that could potentially mitigate this challenge is to create a legal entity in the EU/EEA that receives data transfers according to the SCCs. This entity can then transfer personal data to recipients in third countries that cannot bind themselves to the SCCs, and these non-EU/EEA recipients can then perform further downstream transfers to additional third parties. This is possible. Other contractual terms that provide similar guarantees to the SCCs can be used to perform such onward transfers.

(*Continued*)

Table 1 (*Continued*)

Regulatory compliance challenge	Proposed solution
Assessing the identifiability of structured individual data	Structured data drawn from an individual's record have the potential to lead to reidentification through a single direct identifier (e.g., name, civic address, or social insurance number) or a combination of multiple indirect identifiers (e.g., age, ethnicity, or profession). To ensure that structured data are subject to appropriate governance, the following measures are recommended: ■ Remove the direct identifiers present in a structured dataset. ■ Categorize the remaining data fields as potential indirect identifiers or presumed nonidentifiers. The former are those that could potentially cause individual reidentification either alone or in combination with other data; the latter are those that do not pose a significant risk of causing individual reidentification either alone or in combination with other data. ■ Once the potential indirect identifiers have been labeled as such, calculate the identifiability of the datasets using quantitative methodologies, such as *k*-anonymization. These quantitative methods should account for the potential indirect identifiers and not the presumed nonidentifiers. The calculated identifiability risk can help determine whether the data should be released in full open access or in a controlled access repository.
Assessing the identifiability of unstructured individual data	Unstructured data drawn from an individual's record sometimes have the potential to lead to individual reidentification. This reidentification occurs where certain features of such unstructured data are compared with other identical or correlated data about that individual from another source to establish a positive match; examples include the reidentification of imaging data, genetic sequence data, or handwritten clinical records through comparison with related external datasets. The reidentification risk of individual-level unstructured data can be assessed as follows: ■ Consult experts to determine which elements of the data pose a risk of enabling individual reidentification. For example, imaging data present reidentification risks that are distinct from those posed by genetic sequence data or handwritten clinical notes. ■ Set up collaborations between domain experts and the intended data stewards to develop a data deidentification and data governance strategy that responds to the unique features of the concerned categories of unstructured data. This collaboration could consist of a combination of organizational controls (e.g., implementing access controls) and deidentification measures (e.g., the removal of characteristics such as part of a genomic sequence or image headers from brain imaging data).
Assessing the identifiability of group-level aggregate data	Structured or unstructured data that are compiled from the record-level data of multiple individuals sometimes have the potential to lead to individual reidentification. This can be done by comparing features of the aggregate data that can reveal the presence or absence of a specific individual's data in the aggregate dataset (referred to in the technical literature as a membership inference attack). Numerous strategies can help safeguard against the risk of such reidentification. For example, the following can help mitigate risks for aggregated data in a structured, quantitative form: ■ Use methods such as differential privacy to assess the risk that the data reveal information that is unique to a single record. ■ Modify the aggregated data through the addition of noise—changes to the aggregated data that hide the contribution of individuals' data to the aggregated output, protecting their privacy. For aggregated data that are recorded in an unstructured, nonquantitative form, similar but slightly different methodologies can be used to assess and mitigate the risk of individual reidentification.

Abbreviations: EEA, European Economic Area; EU, European Union; GDPR, General Data Protection Regulation; SCC, standard contractual clause.

A first horizon scan was undertaken to identify the range of ethical issues HCA researchers would face, based on tissue type, sampling scenarios, and donor communities, resulting in a white paper titled "Building the Human Cell Atlas: Issues With Tissues" (4). This paper examines and discusses national legal and ethical norms from seven countries (representative of HCA regions), with specific attention to the similarities and differences across countries and tissue types (gametes, living donors, tissues from deceased individuals, etc.). The paper's aim is to familiarize researchers working with human tissue with the central ethical and legal issues regarding acquisition and use, including consent models, ownership and control of tissues, secondary use, privacy considerations, and so on.

The HCA Ethics Toolkit includes a series of general tools that aim to explain the HCA project and its governance in simple terms, providing useful background documents for institutional ethics review committees tasked with approving collection of tissues and contribution of data to the HCA. Matters addressed in these documents include a simple description of the HCA consortium and its scientific aims, networks, organizational governance, and guiding ethical principles. Different tissue-sampling scenarios are described, and guidance on minimal consent requirements is provided, to guide contributors in adapting their local recruitment procedures and consent tools. A nontechnical description of the HCA DCP is provided, with an overview of ethical requirements to contribute datasets, as well as different access tiers. Finally, in light of the complex international regulatory landscape, particularly with respect to data protection, explanations are provided regarding contributing to the HCA in a way that is compliant with such requirements.

The consent tools propose consent form templates for different HCA sampling scenarios, including a consent template for adult participants, an addendum to consent forms for sampling of clinical leftover tissues, a consent template for deceased donors, and templates for the collection of developmental tissue samples. These templates are based on the identification of a series of core consent elements, presented in **Table 2**, which are considered to constitute appropriate consent from tissue donors to enable open access through the HCA DCP.

With respect to the inclusion of legacy tissue samples and data (i.e., tissues collected before the creation of the HCA), a legacy consent assessment filter tool was created, providing a stepwise questionnaire to screen existing consent forms for use of samples and depositing of data with the HCA (see overview of steps in **Figure 1**). This tool is based on the core consent elements identified above and provides potential steps to take should the contributor's samples or data not have adequate consent conditions (obtaining a waiver from a local ethics committee, reconsenting

Table 2 Core consent elements for depositing data in the Human Cell Atlas public (open) access tier

Core consent element	Consent should be obtained for. . .
Research data	Genetic analysis of data from the tissue sample and the collection of metadata related to the sample
International sharing	International sharing of data
Future use	Any future unspecified use of data
Commercial use	Use of data for commercial purposes
Public (open) access	No access controls or tracking of who is able to access data or for what use
Storage on cloud servers	Storage of data outside the country where they were collected (i.e., the consent language does not restrict storage of data on cloud servers, including private/commercial cloud service providers)
Duration of storage	Indefinite data storage
Data withdrawal	The fact that it is not possible to withdraw data that has already been distributed and used
Reidentification	Risk that the participant could be reidentified in the future, including through linkage with external databases

Table adapted with permission from Reference 30.

Step 1: source of tissue sample

- Does the provenance of the tissue allow for the proposed use of genetic data? (For example: Authorization was obtained to generate genetic data from a legacy sample, the sample was obtained from a deceased donor under a legal framework that does not require ethics review for future research, or data were derived from a tissue that is considered anonymous and permission was obtained to deposit the data open access.)

Step 2: donor consent to broad, international data sharing

- Are appropriate consent elements in place to enable broad, international data sharing? (This includes consent for genetic analyses, international sharing, storage outside the original jurisdiction, indefinite retention, no full withdrawal possible, risk of reidentification, and commercial use.)

Step 3: donor consent for public (open access) sharing

- Does the consent form present language indicating that access to data will be managed or controlled in any way?
- Does the consent form allow for future unspecified use?
- Does the consent form explain that data will be shared through a public (open access) database, meaning that there are no access controls or tracking of who is able to access data or for what use?

Step 4: reconsent of donors or seeking a waiver of consent

- Does the consent form allow for recontact/reconsent of donors?
- Is recontact/reconsent feasible?

Figure 1

Overview of the Human Cell Atlas retrospective consent assessment tool. This figure gives examples of the types of elements examined as part of the assessment tool; please refer to the full tool for an exhaustive description of the process (**https://humancellatlas.org/ethics**).

participants, etc.). This type of tool can be particularly useful to identify the different layers of considerations prior to the use of legacy samples and datasets, particularly in the context of open sharing (52).

Tools for sample and data sharing between sites have also been developed, as it is anticipated that different data contribution scenarios may arise in the HCA. For instance, there may be cases where the research group or institution collecting tissue samples needs to transfer these tissues to another research group or institution in order to undertake molecular analysis (including sequencing). Therefore, to enable downstream contribution of data to the HCA, a template material/data transfer agreement was developed to provide examples of contractual clauses that can be customized to such situations. In particular, these documents provide examples of language used to limit restrictions on use of research materials, as ultimately such restrictions could impact whether data derived from the samples can be contributed to the HCA. Furthermore, template clauses pertaining to intellectual property are suggested, as drafters of material/data transfer agreements should be vigilant regarding any provisions related to data ownership or licensing that may include restrictions on downstream use that consequently limit the ability of one of the parties to contribute open access data to the HCA.

In addition to the above tools, a pediatric portfolio is available alongside the main HCA Ethics Toolkit to address certain issues specific to the Pediatric Cell Atlas (49). This portfolio includes templates for the assent of minors as well as the consent of mature minors, parents, and legally

authorized representatives. Thematic background primers have also been developed to explore how the 1989 United Nations Educational, Scientific, and Cultural Organization (UNESCO) Convention on the Rights of the Child (50) enshrines the right of the child to "the highest attainable standard of health" (article 24) and contributes to the matters of freedom to conduct scientific research, the right to benefit from scientific advances, and the right to nondiscrimination, which together provide the foundations for ongoing and future pediatric research. A primer on data protection and privacy issues related to pediatric research has been developed that explores themes such as regulation of pediatric data protection and data sharing, including consent to data sharing under data protection laws compared with consent/assent to data sharing in the context of the application of research ethics requirements, data retention, the right to object to data processing, and data erasure under the GDPR. Indeed, while there are unique privacy considerations with pediatric data sharing, restricting data sharing can impede certain research on pediatric conditions, which are often distinct from those present in the adult population (40).

Finally, to assist the research community with the implementation and identification of appropriate resources, the HCA Ethics Helpdesk was set up and actively receives queries. This resource helps ensure that resources remain relevant and useful to HCA contributors as the consortium grows and the atlas becomes more refined.

5. CONCLUSION AND FUTURE DIRECTIONS

If the recent global health crises have taught us anything, it is that appropriate data collection and efficient data sharing are crucial to a coordinated healthcare response. The matter of international data sharing has been at the forefront of genomic research for years (31) because no single team can collect, store, and organize the amount of data required to power genomic analysis in studies of common and rare diseases. However, recent reforms in data protection and privacy regulations around the world, such as the GDPR, have made it complex for the scientific research community to understand how health-related data are impacted and how data-sharing consortia must adjust their practices. This is particularly true with respect to open models of data sharing.

Nonetheless, these shifts in how data protection is regulated have also fostered the use of different approaches to data management, including anonymization techniques, contracts (or contractual clauses), sector-specific codes of conduct, data protection impact assessments, and so on. In addition, while there continues to be a certain degree of confusion regarding the interplay of research ethics and data protection requirements (for instance, on the matter of consent), the HCA experience has shown that the development of appropriate ethics governance frameworks and implementation tools can actually serve to complement data protection approaches in an open science context. These governance approaches allow for a reflection on how to implement open science throughout the data ecosystem and adopt proportionate approaches—from research consent (appropriate language for open data sharing) through data contribution (determining appropriate data tiers to implement based on sensitivity of data/metadata and applying anonymization techniques) and data sharing (ensuring appropriate levels of openness and oversight where required). Indeed, recognizing the importance of encouraging the sharing and use of publicly funded data for scientific research purposes, the recently enacted EU Data Governance Act recognizes the importance of adopting appropriate governance mechanisms "in accordance with the principle of being 'as open as possible and as closed as necessary'" (20, recital 16).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

B.M.K., A.B., and E.K. are funded by the Chan Zuckerberg Initiative, the Klarman Family Foundation, and a grant from the Leona M. and Harry B. Helmsley Charitable Trust to McGill University.

LITERATURE CITED

1. 1000 Genomes Proj. Consort. 2015. A global reference for human genetic variation. *Nature* 526:68–74
2. ALLEA (All Eur. Acad.), EASAC (Eur. Acad. Sci. Advis. Counc.), FEAM (Fed. Eur. Acad. Med.). 2021. *International sharing of personal health data for research*. Rep., ALLEA, Berlin, Ger.; EASAC, Halle, Ger.; FEAM, Brussels, Belg.
3. Bahlai C, Bartlett LJ, Burgio KR, Fournier AMV, Keiser CN, et al. 2019. Open science isn't always open to all scientists. *American Scientist*, March–April, pp. 78–82. **https://www.americanscientist.org/article/open-science-isnt-always-open-to-all-scientists**
4. Beauvais MJS, Kirby E, Knoppers BM. *Building the Human Cell Atlas: issues with tissues*. White Pap., Human Cell Atlas. **https://drive.google.com/file/d/1rkUZ9XP2Gs_DzIEH_BoUqq4xU4ZnIgpn**
5. Beauvais MJS, Knoppers BM, Illes J. 2021. A marathon, not a sprint – neuroimaging, Open Science and ethics. *NeuroImage* 236:118041
6. Bernier A, Knoppers BM. 2021. Biomedical data identifiability in Canada and the European Union: from risk qualification to risk quantification? *SCRIPTed* 18:4–56
7. Bernier A, Molnár-Gábor F, Knoppers BM. 2022. The international data governance landscape. *J. Law Biosci.* 9:lsac005
8. Bovenberg J, Peloquin D, Bierer B, Barnes M, Knoppers BM. 2020. How to fix the GDPR's frustration of global biomedical research. *Science* 370:40–42
9. Court Justice Eur. Union. 2016. *Patrick Breyer v. Bundesrepublik Deutschland*. Doc. 62014CJ0582, Case C-582/14, ECLI ID ECLI:EU:C:2016:779
10. Court Justice Eur. Union. 2020. *Data Protection Commissioner v. Facebook Ireland Limited and Maximillian Schrems*. Doc. 62018CJ0311, Case C-311/18, ECLI ID ECLI:EU:C:2020:559
11. Dankar FK, El Emam K. 2013. Practicing differential privacy in health care: a review. *Trans. Data Priv.* 6:35–67
12. Dyke SOM. 2020. Genomic data access policy models. In *Responsible Genomic Data Sharing: Challenges and Approaches*, ed. X Jiang, H Tang, pp. 19–32. London: Academic
13. El Emam K, Arbuckle L. 2014. *Anonymizing Health Data: Case Studies and Methods to Get You Started*. Sebastopol, CA: O'Reilly Media
14. Eur. Comm. 2021. *Commission implementing Decision (EU) 2021/914 of 4 June 2021 on standard contractual clauses for the transfer of personal data to third countries pursuant to Regulation (EU) 2016/679 of the European Parliament and of the Council*. O.J. L 199, June 6, pp. 31–61
15. Eur. Data Prot. Board. 2019. *Opinion 3/2019 concerning the questions and answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection regulation (GDPR)*. Opin. 3/2019, Eur. Data Prot. Board, Brussels, Belg.
16. Eur. Data Prot. Board. 2020. *Recommendations 01/2020 on measures that supplement transfer tools to ensure compliance with the EU level of protection of personal data*. Recomm. 01/2020, Eur. Data Prot. Board, Brussels, Belg.
17. Eur. Med. Agency. 2019. *European Medicines Agency policy on publication of clinical data for medicinal products for human use*. Policy EMA/144064/2019, Eur. Med. Agency, Amsterdam, Neth.
18. Eur. Parliam. 1995. *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. O.J. L 281, Nov. 23, pp. 31–50
19. Eur. Parliam. 2016. *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 (General Data Protection Regulation)*. O.J. L 119, May 4, pp. 1–88. Corrigendum. 2018. O.J. L 127, May 23, pp. 2–5

20. Eur. Parliam. 2022. *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)*. O.J. L 152, June 3, pp. 1–44
21. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, et al. 2003. The International HapMap Project. *Nature* 426:789–96
22. González DR, Carpenter T, van Hemert JI, Wardlaw J. 2010. An open source toolkit for medical imaging de-identification. *Eur. Radiol.* 20:1896–904
23. Granados Moreno P, Ali-Khan SE, Capps B, Caulfield T, Chalaud D, et al. 2019. Open science precision medicine in Canada: points to consider. *FACETS* 4:1–19
24. Greenleaf G. 2021. *Global data privacy laws 2021: despite COVID delays, 145 laws show GDPR dominance*. Rep., Priv. Laws Bus., Middlesex, UK
25. Gunst S, De Ville F. 2021. The Brussels effect: how the GDPR conquered Silicon Valley. *Eur. Foreign Aff. Rev.* 26:437–58
26. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. *Science* 339:321–24
27. Hallinan D, Bernier A, Cambon-Thomsen A, Crawley FP, Dimitrova D, et al. 2021. International transfers of personal data for health research following Schrems II: a problem in need of a solution. *Eur. J. Hum. Genet.* 29:1502–9
28. Haniffa M, Taylor D, Linnarsson S, Aronow BJ, Bader GD, et al. 2021. A roadmap for the Human Developmental Cell Atlas. *Nature* 597:196–205
29. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genet.* 4:e1000167
30. Hum. Cell Atlas. 2020. *Core Research Consent Elements for public (open) data sharing*. Guid. Doc., Hum. Cell Atlas. **https://drive.google.com/file/d/1z-mTEtVGg_ZKg-D6wxOf0mL0BQ9JfpjL/view**
31. Kosseim P, Dove ES, Baggaley C, Meslin EM, Cate FH, et al. 2014. Building a data sharing model for global genomic research. *Genome Biol.* 15:430
32. Kuner C, Bygrave LA, Docksey C, Drechsler L, Tosoni L, eds. 2021. *The EU General Data Protection Regulation: A Commentary; Update of Selected Articles*. Oxford, UK: Oxford Univ. Press
33. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
34. Lindeboom RGH, Regev A, Teichmann SA. 2021. Towards a Human Cell Atlas: taking notes from the past. *Trends Genet.* 37:625–30
35. Liss J, Peloquin D, Barnes M, Bierer BE. 2021. Demystifying *Schrems II* for the cross-border transfer of clinical research data. *J. Law Biosci.* 8:lsab032
36. Majumder PP, Mhlanga MM, Shalek AK. 2020. The Human Cell Atlas and equity: lessons learned. *Nat. Med.* 26:1509–11
37. Majumder PP, Mhlanga MM, Shalek AK, Guigó R, Knoppers BM, Wold B. 2022. How to ensure the Human Cell Atlas benefits humanity. *Nature* 605:30
38. OECD (Organ. Econ. Co-op. Dev.). 2013. *Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data*. Doc. OECD/LEGAL/0188, OECD, Paris
39. OECD (Organ. Econ. Co-op. Dev.). 2016. *Recommendation of the council on health data governance*. Doc. OECD/LEGAL/0433, OECD, Paris
40. Patrinos D, Knoppers BM, Laplante DP, Rahbari N, Wazana A. 2022. Sharing and safeguarding pediatric data. *Front. Genet.* 13:872586
41. Purtova N. 2018. The law of everything. Broad concept of personal data and future of EU data protection law. *Law Innov. Technol.* 10:40–81
42. Raisaro JL, Tramèr F, Ji Z, Bu D, Zhao Y, et al. 2017. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. *J. Am. Med. Inform. Assoc.* 24:799–805
43. Rambla J, Baudis M, Ariosa R, Beck T, Fromont LA, et al. 2022. Beacon v2 and Beacon networks: a "lingua franca" for federated data discovery in biomedical genomics, and beyond. *Hum. Mutat.* 43:791–99
44. Regev A, Teichmann S, Rozenblatt-Rosen O, Stubbington M, Ardlie K, et al. 2018. The Human Cell Atlas white paper. arXiv:1810.05192 [q-bio.TO]

45. Rozenblatt-Rosen O, Shin JW, Rood JE, Hupalowska A, Regev A, Heyn H. 2021. Building a high-quality Human Cell Atlas. *Nat. Biotechnol.* 39:149–53
46. Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. 2017. The Human Cell Atlas: from vision to reality. *Nature* 550:451–53
47. Shabani M, Borry P. 2018. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. *Eur. J. Hum. Genet.* 26:149–56
48. Shabani M, Marelli L. 2019. 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.* 20:e48316
49. Taylor DM, Aronow BJ, Tan K, Bernt K, Salomonis N, et al. 2019. The Pediatric Cell Atlas: defining the growth phase of human development at single-cell resolution. *Dev. Cell* 49:10–29
50. UNESCO (UN Educ. Sci. Cult. Organ.). 1995. *The Convention on the Rights of the Child: UNESCO's contribution*. Progr. Meet. Doc., UNESCO, Paris
51. Vayena E, Gasser U. 2016. Between openness and privacy in genomics. *PLOS Med.* 13:e1001937
52. Wallace SE, Kirby E, Knoppers BM. 2020. How can we not waste legacy genomic research data? *Front. Genet.* 11:446
53. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, et al. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 604:437–46