E.8.1 Data Access Framework

In the initial stages of ICGC ARGO we are adopting the existing ICGC Data Access Policy, published December 2012. This policy is now under revision and will be released in due course.

The nature of the data that will be produced by ICGC-ARGO members; substantial clinical annotation and extensive genomic data, raises important human subject privacy protection issues. The patient/individual protection policies developed for ICGC-ARGO are designed to balance two important goals: to facilitate investigations of genomic changes related to cancer and, at the same time, to respect and protect the patients/individuals whose data and materials have been or will contribute to ICGC-ARGO member programs. It is technically possible that genomic information generated by ICGC-ARGO could lead to re-identification of an individual if linked or combined with other information or archived data There is also a risk of individual identification by computer-based analysis of the clinical data in conjunction with, for example, third-party demographic and healthcare management databases. This potential identification could then publicly link the individual to his/her clinical information collected by the participating projects and could lead to social risks such as discrimination or loss of privacy.

ICGC ARGO member programs will have privileged access to data from other members of the Consortium based on their level of Membership. After a 24-month period following standardized analysis ICGC ARGO data will be made available to external parties following established data access processes described below. Data users will be required to consult the ICGC ARGO Publication Policy to be aware of the publication status of data sets and guidelines in place on behalf of data producers.

ICGC-ARGO have carefully considered, based on existing knowledge and best practice, which data types should be publicly accessible, and which should be governed by a controlled process.

POLICY: To minimize the risk of patient/individual identification, the ICGC has established the policy that datasets be organized into two categories, open and controlled access. Table 1 includes a list of data elements and the data access category within which they will be available.

The first category, Open Access Datasets, will be publicly accessible and contain only data that cannot, at present, be aggregated to generate a dataset unique to an individual without reasonable efforts.[1] The amount and nature of genetic data that might be associated with an individual from the Open Access Datasets has been carefully considered and will continue to be monitored by ICGC. The second category, Controlled Access Datasets, will contain composite genomic and clinical data that are associated to a unique, but not directly identified, person.

ICGC Open Access Datasets	Controlled Access Datasets
Cancer pathology · Histologic type or subtype · Histologic nuclear grade · Tumor staging **Patient/person** · Gender · Age (single category for ages over 89) · Vital status · Age at last follow-up (single · category for ages over 89) · Survival time · Cause of death · Relapse type · Relapse interval · Disease status at last follow-up · Interval from primary diagnosis to last follow-up · Treatment type · Treatment duration · Therapeutic intent · Response to therapy · Cumulative drug dosage **Specimen/sample** · Specimen tissue source · Specimen anatomic location **Molecular Data** ● Gene expression (normalized) ● DNA methylation ● RNA-Seq read counts (unnormalized) • Genotype frequencies • Computed copy numbers and loss of heterozygosity · Newly discovered somatic variants	**Detailed Phenotype, treatment and outcome data** • Region of residence • Ethnicity • Risk factors • Post therapy staging • Performance status • Detailed treatment cycle and dose details • Treatment toxicity **Specimen/sample** • Specimen processing **Molecular Data** ● Gene Expression (probe-level data) ● Raw genotype calls ● Gene-sample identifier links ● Genome sequence files

Table 1. Listing of data categories and level of access restriction on those data.

This list will be periodically revised to reflect the continually evolving fields of genomics, bioinformatics, and to comply with ethics and privacy policies and regulations.

ICGC established two bodies to oversee controlled access: The Data Access Compliance Office (DACO) and an International Data Access Committee (IDAC). DACO is responsible for processing access requests from the scientific community and its activities are overseen by IDAC. DACO is required to verify the conformity of users' projects with the goals and policies of ICGC, including, but not limited to, policies concerning the purpose and relevance of the research, the protection of participants, and the security of participants' data.

DACO, IDAC, and ICGC's Ethics and Governance Committee (AEGC) collaboratively developed the data access application forms (which include an access agreement), as well as the policies to be used by ICGC. The rules and policies of ICGC have influenced the controlled access strategies of several database projects, including the Wellcome Trust Sanger Institute and the Human Epigenome Consortium.

Authorizations to access controlled data will be broad, so that authenticated users will get permission to obtain access to controlled data generated from all samples studied by any participating ICGC ARGO project (as the feasibility of providing permissions to datasets originating from single or partial subsets of participating center's has been determined to be unworkable in the context of the ICGC).

The DACO will also develop guidelines to streamline approaches to providing qualified investigators with access to controlled data. In doing so, it will consider mechanisms and tools that have been already in use by other organizations that distribute controlled datasets to international scientists (for example, GA4GH or the Wellcome Trust Case Control Consortium). Under current processes potential users and their institutions will be required to submit an Access Application Form and sign a Data Access Agreement. Interested users and institutional officials who are authorized to make legally binding agreements for the institution will be required to adhere to the conditions laid out in the Access Agreement. Investigators will need to agree to regular review and renewal requested by the DACO for such authorization and in cases when they move to new institutions.

[1]Council of Europe, *Recommendation Rec (2006)4 of the Committee of Ministers to member states on research on biological materials of human origin*

Appendix I: Resource of Applicable ICGC ARGO Member Program Laws and Regulatory Frameworks Guiding Privacy, Data Protection and Sharing

Appendix II: Further Reading
Ethical and Legal Codes and Policies Guiding Data Sharing Behavior

References:

Policy Section E.8.1
Version 3.0 Published February 2021

[1] Zornita Stark et al: Integrating Genomics into healthcare: A Global Responsibility. American Journal of *Human Genetics*, 104, 13-20, January 3 2019.

[2] 1948 Declaration of Human Rights (art. 27)

[3] OECD Recommendation on health data governance in 2017

[4] UNESCO Science and Scientific Researchers Guidelines 2017 (ref UNESCO 2017).

[5] NIH Genomic Data Sharing Policy: https://osp.od.nih.gov/wp-content/uploads/NIH_GDS_Policy.pdf

[6] Prepublication data sharing, Toronto International Data Release Workshop Authors. *Nature* **461**, 168–170 (2009).

[7] Jane Kaye, Data sharing in genomics — re-shaping scientific practice. *Nature Genetics,* May 2009.

[8] Yann Joly et al: Analysis of five years of controlled access and data sharing compliance at the International Cancer Genome Consortium, *Nature Genetics*. 2016 Mar;48(3):224-5.

[9] GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data. https://www.ga4gh.org/wp-content/uploads/Framework-Version-10September2014.pdf. Accessed November 2020.

[10] Wilkinson, M. D.et al.The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 (2016).