Integrated Canine Data Commons
Data Governance Advisory Board
Draft Guidelines for Evaluating Data Submissions During Prototype Phase

August 2019

Abbreviations

Term	Meaning
CRDC	Cancer Research Data Commons – a network of nodes brought together by the NCI to share cancer related data
DCF	Data Commons Framework – reusable framework that provides Authentication and Authorization and indexing services
IACUC	Institutional Animal Care and Use Committee
ICDC	Integrated Canine Data Commons - part of the CRDC that contains canine study data for broad sharing
IRB	Institutional Review Board
DGAB	Data Governance Advisory Board – this body which evaluates submissions for inclusion in the ICDC
FFRDC	Federally Funded Research and Development Center
FNL	Frederick National Laboratory – an FFRDC that facilitates cancer research on behalf of NCI
NCI	National Cancer Institute of the NIH
NIH	National Institutes of Health

Introduction:

The Integrated Canine Data Commons (ICDC) will receive data from many projects and provide the community with relevant access to that data. During the prototype phase, not all requests to upload data to the ICDC can be accommodated due to the ICDC's focus and the effort and costs associated with bringing the data into the ICDC. Therefore, the ICDC must develop, document, and adhere to a Data Governance Process. This Data Governance Advisory Board will define and administer that process.

DGAB Mission:

Advise NCI Senior Advisory Committee, based on measurable criteria, on the priority for input into the ICDC of various data sets as received from potential submitters. Devise and publish a submission process for potential submitters. Publish results of submission evaluations and priority to ensure transparency. Report to the Steering Committee on a regular basis.

Process:

1. Initiation of submission request – submitter completes submission packet for data set. Submissions are accepted on a scheduled basis to be reviewed by the DGAB. Any

submissions made after the deadline for that submission period will be reviewed on the next submission period.

2. FNL submission packet review – FNL staff will review the submission packet for completeness and provide feedback to the submitter as needed. Only completed packets will be forwarded to the DGAB for evaluation.

3. DGAB evaluation – On a regular basis, the DGAB will meet virtually and review, evaluate, and prioritize the submission packet. Prioritization will be done across the entire portfolio of submissions and will be updated as new submissions are received.

4. DGAB recommendation – Following the DGAB meeting, a recommendation on priorities will be made to the NCI Senior Advisory Committee.

5. NCI Senior Advisory Committee recommendation – The NCI Senior Advisory Committee will determine final priority of submissions to the ICDC. Priority determinations will be published on the ICDC website. Based on these recommendations, the ICDC Data Management Team will begin working with the submitters to prepare their data for entry into the ICDC.

Request Initiation.
A web-based form will be used to collect initial information. The following information will be collected from the submitter:

1. Name/Identifier of Study
2. Attach Study Protocol
3. Grant ID and funding source (if applicable)
4. Scientific Point of Contact (Name, Phone, Email)
5. Data Manager Point of Contact (Name, Phone, Email)
6. Attach current IACUC/IRB approval documentation for this study (if applicable)
7. Data access policy (choose one): Open-access – no-embargo, Controlled-access – no embargo, Open-access – embargo, Controlled-access - embargo
8. Cancer type(s) included in study
9. Number of subjects included in study
10. Data types included in study (check all that apply): Imaging, genomics, proteomics, immunology, clinical, other (specify)
11. Approximately how much data (in TB) do you have?
12. In your own words, describe the overall scientific benefit of including this study in the ICDC prototype.
13. List any publications associated with this study, if any.
14. Are there any time constraints on processing/loading/releasing the data?
15. Attach Data Dictionary
16. Attach Data Model/Schema diagram indicating how collected data relates to subjects, visits, samples, etc.

17. Have you mapped any of your data to a standard such as SEND? If so, which standard was used?
18. Anticipated budget needed to prepare data set for submission.

Evaluation Criteria:
Studies will be evaluated on the clarity of the biological question being asked, the quality and quantity of the data obtained and the perceived value of the data set with regards to further explaining/treating cancer in humans and canines. In particular, the following will be evaluated:

1. Clear statement of clinical or research intent for the study.
2. Impact on cancer research.
3. Scientific approach.
4. Focus on innovation/discovery in cancer research.
5. Some factors which may contribute to the selection process are:
    a. Characterization of disease, pathology, histology with standardized, mappable nomenclature.
    b. Longitudinal measurement of disease, disease progression, treatment and response
    c. Molecular measurements (epigenetic, whole genome/exome sequencing, RNA-seq, tumor, as applicable), including:
        i. instrument and sample prep and handling protocols,
        ii. analysis pipeline and versioning,
        iii. depth and quality of sequence,
        iv. mapping and variant calling
    d. Proteomic measurements
    e. Immunology and immune system characterization (including immunohistochemistry slides)
    f. Metabolomic measurements
    g. Microbiome data
    h. Single cell data
    i. Spatial and molecular measurements, such as nanostring
    j. Imaging (ultrasound, MRI, CT, cryoEM, etc.)
    k. Characterization of individual tumors with cell culture, 3D cell culture, PDX, Organoid
    l. Data availability to the public

The ultimate goal of the DGAB is to choose studies that will aid the ICDC in creating a ‘Rosetta stone’ for mapping disease, data types, interventions, and response to human disease, so that the analysis of that data will identify cancers that are likely to have similar/different biological drivers and thus respond or not respond to a given therapy similarly between canines and humans. These studies are likely to elucidate novel and hopefully informative mechanisms. The drivers for the ICDC are to 1. Create a high value resource that will aid investigators in primary and secondary analysis. 2. Allow investigators to compare their data with other canine and comparative datasets. 3. Be a trusted, high quality, highly accessed, highly shared resource for canine oncology datasets. 4. Encourage collaborative research and provide tools for the canine oncology community to collaborate on problems that are important in comparative oncology.

Prioritization:

Preference is given to data sets which can be fully public and do not require any application process or data use agreements.

The DGAB will evaluate each study individually on its merits and will then assign a priority to the study relative to studies previously evaluated. In theory, this means that a particularly valuable study, as determined by the DGAB evaluation, could interrupt the resource allocation to existing studies "in-process" and be given a higher priority. In practice, it is likely that studies will be allocated resources based on submission timing.

ICDC-DGAB Evaluation Criteria Detail:

Data may be submitted to the ICDC that span the breadth of clinical, pathologic and -omics studies, aimed at advancing our collective knowledge of cancer in humans and dogs. While all data sets submitted must be associated with minimum requirements for common data elements (CDEs), a subset of data sets will be selected for ICDC assistance to transform the data into the optimum format needed to maximize their impact. The selection of data sets will primarily be based of identifying studies where there is a clear biological question being asked, the quality and quantity of the data provided is high, and there is strong perceived value of the data set to align with the goals of the ICDC.

The information provided below will be used by the DGAB to help determine which data sets would be most suited to assist in processing through the platform.

See ICDC Data Guidelines (below) for minimum expected data.

A) Basic Study Information

1) Number of animals involved
2) Number of samples per animal
3) Narrative for inclusion/importance of study in ICDC
4) Experimental protocols (attach here)

B) Clinical data:

1) Data standard used in data collection (e.g., SEND, CDISC, BRIDG, caDSR CDEs, etc.)
2) Describe completeness of minimum common data elements (CDEs)
3) Sample type(s)
4) Does the study include outcome data (yes/no)?
5) Clinical data dictionary (attach here)

C) Biospecimen data:

1. Source of the biological specimen(s) used to isolate DNA (Check all that apply):
   - □ Blood - select one or more of EDTA/ACD/PaxGene
   - □ Tumor tissue – state anatomical location:
   - □ Non-tumor tissue - state anatomical location:
   - □ Fresh frozen
   - □ Fixed – formalin type and duration:
   - □ Cell line
   - □ Fresh
2. Indicate organization of longitudinal samples and number of biological replicates and/or technical replicates (if any)
3. Attach biospecimen protocol and data dictionary

D) DNA data:

1) Experimental approach used in the study (Check all that apply):
   - □ Whole genome sequence
   - □ Whole exome sequence
   - □ Targeted genomic DNA sequence
   - □ Other: □
2) Method of DNA quality control? (Check all that apply):
   - □ UV/VIS
   - □ Bioanalyzer
   - □ TapeStation
   - □ AGE
   - □ Other: □
3) Attach DNA protocol and data dictionary

E) Transcriptomic data:

1) Specify experimental approach used:
   - □ Whole transcriptome RNA-seq
   - □ mRNA-seq
   - □ small RNA profiling
   - □ miRNA-seq
   - □ scRNA-seq
   - □ Other: □
2) QC methods applied (Bioanalyzer, RIN >7.0, etc.)

   □
3) Attach transcriptomic protocol and data dictionary

F) Epigenomic data:

1) Specify experimental approach used:
   - □ ChIP-Seq (specify antibody target)
   - □ ATAC-seq
   - □ HiC-seq
   - □ methyl-seq
   - □ Bisulfite-seq
   - □ Other: □
2) QC methods applied:

   □
3) Attach epigenomic protocol and data dictionary

G) Proteomic data

1) Specify experimental approach used:
   - □ LC-MS
   - □ Targeted assay (SRM)
   - □ Protein Arrays
   - □ Other: □
2) QC methods applied:

3) Type of post-translational modification characterization (Glycosylation, phosphorylation, Acetylation, Methylation, etc.) (if any):

4) Attach proteomic protocol and data dictionary

H) Imaging data:

1) Specify imaging modality used:

□ Ultrasound
□ MRI
□ CT
□ X-ray
□ PET
□ Whole Slide Imaging
□ H&E
□ cryoEM
□ Other:

2) Imaging standard used (if applicable)

3) Quality control metrics and curation process

4) How does the imaging relate to the specimens acquired?

I) Characterization of individual tumors with cell culture, 3D cell culture, PDX, Organoid

J) Data availability to the community

The ultimate goal of the ICDC is to select studies that will support the development of robust and accessible canine data sets that are likely to elucidate novel and hopefully informative mechanisms.

The data selected for inclusion are intended to aid with mapping canine disease, data types, interventions, so that their analyses will identify cancers likely to have similar/different biological drivers to the corresponding human cancers. This approach will inform the comparability of response to a given therapy between canines and humans.

Are there any limits to making this data available to the community and, if so, what are they?

ICDC Data Guidelines

It is expected that most data will be organized into nodes similar to:
Studies→Cases→Evaluations→Samples→Lab Data→Files

The information presented here represents the minimum data expected for each of these nodes.

Studies:

Data Element	Description	Priority
study name	A one sentence title that will be used in the display of Study records within the UI	Required
study description	A short (3-6 sentence) summary of the principal aims of the study> This will be displayed within the UI as a key part of the Study "details" view	Required
study code	An alpha-numerical ID by which the study can be uniquely identified once loaded into the ICDC. This will appear prominently within the UI and should therefore be "human-friendly". This will also be used to generate globally unique Case IDs from patient/subject/donor IDs that may only be unique within any given study.	Required
dates of conduct	Approximate dates upon which the study started and ended, such that study "detail" view can clearly communicate when and for how long the study was conducted	Preferred
date of IACUC approval	If applicable, providing the date of the appropriate IACUC approval would add to the "completeness" of Study records, but is probably not something of major importance to consumers of the data	Optional
study arms	For the COTC studies, we can derive information as to the arms represented within a study from the study protocol and/or the "ENROLL_TX_ASSIGN_CD_FUL" field within the C3D Enrollment CRF. Having the study arms explicitly stated by the data provider would, however, enable us to create and name the appropriate study arms exactly as the data owners would like to see them presented within the UI.	Preferred
study cohorts	Likewise, for the COTC studies, we can derive information as to the cohorts represented within any given study from the study protocol and/or the "ENROLL_TX_ASSIGN_CD_FUL" within the C3D Enrollment CRF. Having the treatment cohorts explicitly stated by the data provider would, however, enable us to create and name the appropriate cohorts exactly as the data owners would like to see them presented within the UI.	Preferred
study protocol(s)	Wherever possible, data owners should provide copies of the relevant study protocol or protocols, such that the documents in question can be uploaded into the ICDC, essentially as downloadable "attachments" associated with the corresponding Study records. Combined, the name and description of any given Study, as described above, would provide an overview of the Study, but access to the detail as to exactly how the study was conducted, inclusion/exclusion	Preferred

	criteria, etc. via protocol documentation would be of significant value to data consumers	

Cases:

Data Element	Description	Priority
patient ID	An ID by which the data owner uniquely identifies patients/subjects/donors, at least within the confines of a single study. This external ID will ultimately be concatenated with the relevant study code described below, in order to create an ICDC Case ID that will uniquely identify patients/subjects/donors across all studies.	Required
study code	An alpha-numerical ID by which the study can be uniquely identified once loaded into the ICDC. Although this represents a Study attribute, it will need to be provided alongside Patient ID such that the two IDs can be concatenated together to generate globally unique Case IDs. Unless the two IDs are provided side by side within the same source data file, it won't be possible to perform the required concatenation.	Required
study arm	Although this represents a Study attribute, alongside the Case-centric data, data owners will have to provide sufficient information as to the study arm to which each patient/subject/donor belongs such that the resulting ICDC Cases can be correctly associated with their Study Arms.	Required
cohort	Likewise, although this also represents a Study attribute, alongside the Case-centric data, data owners will have to provide sufficient information as to the dosing cohort to which each patient/subject/donor belongs such that the resulting ICDC Cases can be correctly associated with their Cohorts.	Required
breed	The specific breed of each patient/subject/donor should be provided according to an appropriate controlled vocabulary of acceptable terms, as identified by the Data Governance Advisory Board.	Required
gender	It would be highly preferable for data owners to provide the gender of each patient/subject/donor, as indicated simply by M for male and F for female, exclusive of any indication of the subject having been spayed or neutered.	Required
neutered status indicator	It would be highly preferable for data owners to provide a simple Boolean (Yes or No) indication as to whether the patient/subject/donor has been spayed or neutered, one that is entirely separate from information as to gender.	Required
weight	Data owners should consistently report this in kg. In the case of longitudinal studies/trials, weight would likely be determined at each visit (i.e. occurrence of a physical exam, sample collection, disease evaluation) and reported as part of these visit-based data collections. But in studies that are not longitudinal in nature, "weight" would be the patient's weight at the time of, for example, samples being collected on a one-off basis prior to them being banked for future analysis. In this situation, weight would be reported as part of case-based data collections.	Preferred
diagnosis	A specific diagnosis should be provided for each patient/subject/donor, according to an appropriate controlled	Required

	vocabulary of acceptable terms, as identified by the Data Governance Advisory Board.	
stage of disease	Wherever possible, a specific indication of the stage of disease should be provided for each patient/subject/donor, according to a controlled vocabulary of acceptable terms from an appropriate staging convention, as identified by the Data Governance Advisory Board.	Preferred
disease site	Wherever possible, a specific indication of the primary site of the cancer should be provided for each patient/subject/donor, according to an appropriate controlled vocabulary of acceptable terms, as identified by the Data Governance Advisory Board.	Preferred
date of diagnosis	Wherever possible, a date for the original diagnosis of the cancer should be provided for each patient/subject/donor. This date, when compared to dates of study enrollment and/or sample acquisition would provide context as to how well established the patient's tumor(s) is/are.	Preferred
study site	Indication as to geographical location of the study site at which the patient was enrolled. Provides some context, but of questionable value to data consumers.	Optional
date of informed consent	Potentially useful relative to Date of Diagnosis?	Optional
date of registration	Indication as to when the patient was enrolled into the study in question - potentially useful relative to Date of Diagnosis?	Optional

Evaluations:

Data Element	Description	Priority
patient ID	The ID by which the data owner uniquely identifies the patients/subjects/donors to which evaluations and observations pertain, at least within the confines of a single study. Although this ID represents a Case attribute, it will ultimately be concatenated with the relevant study code described below, in order to create an ICDC Case ID that will uniquely identify the patients/subjects/donors to which evaluations and observations belong, across all studies. This ID must therefore be provided alongside the evaluation and observation data.	Required
study code	The alpha-numerical ID by which the study can be uniquely identified once loaded into the ICDC. Although this represents a Study attribute, it will also need to be provided alongside the evaluation and observation data itself, such that Patient ID and study code can be concatenated together to generate the globally unique Case IDs to which evaluation and observation data belong. Unless these two IDs are provided side by side within the same source file of evaluation and observation data, it won't be possible to perform the required concatenation and subsequent data-to-case mapping.	Required
date	Data owners must provide a date upon which each and every longitudinal "visit-based" evaluation or observation occurred. This date, in association with the globally unique Case ID derived as described above, will be used to create globally unique Visit records to which sets of evaluation and observation data, such as physical exam observations and extent of disease assessments, can be associated. Creating Visit records in this way will allow data consumers to, for example, find and analyze the physical exam observations that were made at the same time as a corresponding set of extent of disease assessments, because the two sets of data will share a visit. And by extension, look for and analyze data from samples taken at those same times.	Required

Samples:

Data Element	Description	Priority
sample ID	The ID by which the data owner uniquely identifies the sample, at least within the confines of a single study. This ID may ultimately be concatenated with the globally unique ICDC Case ID for the patient/subject/donor in order to uniquely identify each sample.	Required
patient ID	The ID by which the data owner uniquely identifies the patients/subjects/donors to which samples belong, at least within the confines of a single study. Although this ID represents a Case attribute, it must be provided alongside the sample data itself such that samples can be mapped to their correct ICDC Case.	Required
study code	The alpha-numerical ID by which the study can be uniquely identified once loaded into the ICDC. Although this represents a Study attribute, it also must be provided alongside the sample data itself, such that Patient ID and study code can be concatenated together to generate the globally unique Case IDs to which samples can be mapped. Unless these two external IDs are provided side by side within the same source file of sample-centric data, it won't be possible to perform the required concatenation and subsequent sample-to-case mapping.	Required
sample collection date	Data owners must provide a date upon which each and every sample was collected. This date, in association with the globally unique Case ID derived as described above, will be used to map each sample to the appropriate Visit for the patient/subject/donor in question. Associating sample collections with visits in this way organizes the sample per se, but also creates a connection between sample acquisitions, and data from corresponding evaluations and observations occurring for the same patient/subject/donor on the same date.	Required
sample type	Indication as to the physical nature of each sample in question, e.g. tissue, whole blood, plasma, etc. It would be highly preferable to specify this information according to an appropriate controlled vocabulary of acceptable terms, as identified by the Data Governance Advisory Board.	Required
sample collection site	Indication as to the anatomical site from which each sample in question was acquired, e.g. lung, skin, lymph node, etc. It would be highly preferable to specify this information according to an appropriate controlled vocabulary of acceptable terms, as identified by the Data Governance Advisory Board.	Required
general sample pathology	Indication as to the pathological nature of each sample in question, e.g. normal, benign tumor, malignant tumor, hyperplasia, etc. If this sample classification attribute is considered to have value, it would be preferable to specify this information according to an appropriate controlled vocabulary of acceptable terms, as identified by the Data Governance Advisory Board.	Preferred
necropsy sample	Indication as to whether the sample in question was acquired during a necropsy, as opposed to having been acquired from a live patient/subject/donor	Optional

Lab Data:

Data Element	Description	Priority
sample ID	The ID by which the data owner uniquely identifies the sample from which the laboratory data has been derived, at least within the confines of a single study. This ID may ultimately be concatenated with the globally unique ICDC Case ID for the patient/subject/donor, in order to uniquely identify each sample.	Required
patient ID	The ID by which the data owner uniquely identifies the patients/subjects/donors to which samples, and thereby the lab data derived from them, belong, at least within the confines of a single study. Although this ID represents a Case attribute, it must be provided alongside both the lab data itself, and the corresponding sample ID, such that the lab data can be mapped to the correct sample and ICDC Case.	Required
study code	The alpha-numerical ID by which the study can be uniquely identified once loaded into the ICDC. Although this represents a Study attribute, it must also be provided alongside the lab data, such that sample ID, patient ID, and study code can be combined in order to map lab data to the correct sample and ICDC Case. Unless these IDs are provided side by side within the same source file of laboratory data, it won't be possible to perform the required concatenation and subsequent data-to-sample-to-case mapping.	Required

Files:

Data Element	Description	Priority
parent type	what type of parent does the file belong to? A diagnosis? A physical exam? A sample? An assay? An aliquot?	
parent record ID		
aliquot ID		
assay ID		
sample ID		
physical exam ID	effectively study+patient+date of exam	
diagnosis ID		
patient ID		
study code		
file name		
file type		
file description		