# National Cancer Institute
# Integrated Canine Data Commons Steering Committee

## Teleconference
## Wednesday, February 20, 2019

---

**Participants**

External Committee Members
Matthew Breen
Renee Chambers
Dawn Duval
Allison Heath
Will Hendricks
Warren Kibbe
Debbie Knapp
Cheryl London
Jeff Trent
Roel Verhaak

Internal Committee Members (NCI and FNL)
Matthew Beyers, ICDC Project Manager
Allen Dearry
Toby Hecht
Amy LeBlanc
Paula Jacobs
Tony Kerlavage
Erika Kwon
Christina Mazcko
Elaine Ostrander
John Otridge
Ralph Parchment, ICDC-SC Managing Secretary
Connie Sommers
Greg Tawa

Others
Mary Cerny (writer)

**Opening and Welcome**

Dr. Parchment opened the meeting at 11:30 a.m., noting that this was the first meeting of the NCI Integrated Canine Data Commons Steering Committee (ICDC-SC). He also noted that the meeting was being recorded to create accurate meeting minutes.

Drs. Parchment, Hecht, and LeBlanc welcomed the Steering Committee members on behalf of the NCI Division of Cancer Treatment and Diagnosis (DCTD), Comparative Oncology Program

(COP), and Frederick National Laboratory for Cancer Research (FNLCR). They thanked the participants for their interest and for contributing their expertise and time to the development of a canine data commons by serving on the ICDC-SC. Through this strategic partnership, the SC will advise NCI on the building of the ICDC prototype and the implementation of the NCI principle of "community-driven" data commons to ensure benefit to the broader research community.

**Introductions**

Committee members introduced themselves and described their background and areas of research interest.

**Orientation and Purpose of the Steering Committee**

In creating and implementing the ICDC, the overarching question for NCI is whether pet dogs with spontaneous cancers can serve as close models of human disease in order to evaluate new drugs, immunotherapeutic agents, and combinations for further development for human cancer patients.

One approach to answering this question is to establish a publicly accessible canine database as an interoperable node in the larger NCI Cancer Research Data Commons (CRDC) that would contain the following:

- Full genotype and phenotype characterization of the major canine tumors and normal tissues, including tumor mutational burden (TMB) and neoantigens (seen by T cells in the context of canine MHC antigens);
- Description of the tumor microenvironment (TME), including numbers and types of subsets of immune (and other) cells; and
- Clinical data from the Comparative Oncology Trials Consortium (COTC) and other canine cancer trials, including images; data from treatment-naïve, post-treatment, and relapsed patients; and all other clinical data elements for canines.

The key NCI focus areas for this project include workforce development, basic science, big data, and clinical trials. The objective of the big data focus area in particular is to increase data aggregation and interpretation in support of work across the cancer enterprise.

Initial funding and resources for the ICDC started in 2016 as part of the Precision Medicine Initiative in Oncology with funding for eight NCI-designated Cancer Centers (P30s) to work with veterinary oncologists and geneticists to:

- Sequence 25 cases of one or more of the following canine tumors and their normal equivalents: B-cell lymphoma, melanoma, glioma, bladder cancer, osteosarcoma, and mammary cancer;
- Determine the TMB and identify neoantigens that can strongly bind canine MHC antigens; and
- Characterize T cell subsets and numbers, as well as other relevant aspects of the TME.

The second NCI initiative, about a year later, was a request for application (RFA) designed to support canine clinical trials using immunotherapeutic agents and novel combinations of immune modulators, molecularly targeted agents, chemotherapy, and/or radiation. This RFA also supports laboratory correlative studies that seek to describe, characterize, and understand the cellular and molecular mechanisms that determine the anti-tumor response (or non-response) in dogs with spontaneous tumors. The infrastructure needed to support this effort includes a network of up to five academic laboratories, veterinary medicine clinical trial sites, and veterinary pharmaceutical companies working together (via UM1s), along with a single coordinating center (U24) assisted by NCI's COP and an NCI program official. The program is funded for 5 years.

The ICDC is one of several data-driven nodes that constitute the larger cloud-based NCI CRDC. Other nodes include proteomics, genomic data commons, imaging, immuno-oncology, cancer models, population studies, clinical studies, and biomarkers. The CRDC has portals for data contributors and consumers to enter the cloud in addition to workspaces, tool repositories, data models and dictionaries, and analytical and validation tools.

The work scope and 5-year timeline of the ICDC is divided into three phases: the ICDC prototype phase (2 years); the ICDC production phase (1 year); and the ICDC production, operation, and maintenance phase (2 years). The prototype phase involves development, implementation, and testing of the prototype; production system cost; and schedule analysis. Once these base requirement tasks are completed, milestone 1 will be reached. The project will then proceed to the production phase, which involves development and implementation of the ICDC (option task 1). Once the prototype system transitions to the fully functioning ICDC, milestone 2 will be reached, and the project will enter the "final" phase of the timeline, in which the ICDC is operational, with ongoing validation, updating, and maintenance (option task 2). Completion of this last phase is designated as milestone 3 (project end).

Major activities undertaken for milestone 1 include harmonizing the data, setting up the ICDC prototype, and establishing and convening the SC. The builders of the prototype include the IT specialists and NCI and FNLCR staff who will manage and support this effort to stand up a prototype as the basic task. The ICDC-SC includes 10 NIH members, 10 members from academia and industry, and 3 members from Leidos Biomedical Research, Inc. The SC will include an End User Working Group (SCWG) composed of research staff at SC institutions. The charge to the SCWG is to compile and curate canine and other datasets with clinical outcomes (if available) for validation exercise and testing and to establish a reference genome considering various breeds. The working group will collect data from the Canine Immunotherapy Trials Network, NCI-designated Cancer Centers, the National Human Genome Research Institute, and other sites. The prototype phase is designed to be interactive, with the builders and the SC working as partners to review and modify design and operations and to define and refine data models and end user tools.

The ICDC is in year 1 of the 2-year prototype phase. The following activities have been accomplished in the first 4 months of this phase of the project:

- Steering Committee: Full membership of the ICDC-SC has been finalized, subcontracts are being established, and the inaugural SC meeting has convened.

- Data: A first data model has been chosen, based on a COTC trial, and will be tested. Receipt of genomic, sequence, and proteomic data from the National Center for Advancing Translational Sciences (NCATS) has started, establishing import mechanisms. Eventually, users will submit their own data. Other researchers have relevant data and are willing to contribute to the ICDC. Use cases are being gathered to determine how the data commons will be used and analyzed.
- Software: Configuration of the ICDC infrastructure and the data model to work within a modified Gen3 software system is underway. NCI CloudOne (hosted by Amazon Web Service (AWS)) was selected as the prototype's cloud infrastructure. The NIH team is meeting with the cloud resources team to understand establishment of interoperability and continues to work with the semantics team.

ICDC prototype development is fully funded and should be complete by September 2020. Parts of option task 1, the production phase, are expected to start by September 2019. New funds from NCI to complete milestones 2 and 3 are being requested due to the fast pace needed to reach milestone 1.

Questions and discussion

Further clarification regarding the projected timeline and phases was requested, specifically as related to the apparent overlap between the ICDC production and ICDC prototype phases. It was noted that the prototype phase has different parts and activities. Under the production phase, new data will be received and imported to actively build the database. To expedite this process, the team does not want to wait for completion of the prototype phase, so building of the database will not be limited to data "on hand" during the prototype phase of the project.

Another question involved how standardization and characterization of TMB across teams will be addressed and achieved for the ICDC, given that investigators use different approaches to measure this parameter. The plan is not only to use TMB as provided by researchers but also to take into account how TMB was derived and then compare findings to determine how best to standardize TMB. There is a uniform, standardized pipeline to process human data, from the raw data to determine the final metric, with appropriate post hoc filtering. This approach is being used with the Genomic Data Commons (GDC), which could serve as a model for standardization of TMB for the ICDC. Having a standardized platform to calculate TMB from whole-genome vs. whole-exome sequences and other panels will also be important for this project. This platform can be built using currently available samples and genomic data, with further contribution from dozens to hundreds of samples that could be added to the repository.

Additional information was requested about follow-up to the activities by NCI-designated Cancer Centers (P30). Each site sequenced 25 cases of the specified canine tumors and their normal equivalents, which would translate into approximately 200 canine tumor sequences. This is a relatively small number when compared to the GDC, which has thousands of tumor sequences. Several projects and plans are in place to build the ICDC database beyond the work done under the supplements. The RFA described above involves testing of immunotherapeutic agents in canine clinical trials. Data will be collected from other ongoing canine trials involving primary tumors, metastases, and testing of pre- and post-treatment sequencing. It was noted that some NCI-designated Cancer Centers contributed more than 25 cases. Although the ICDC is not

yet at the scale needed for a robust commons, submissions are expected to increase going forward as the project progresses. For example, NCATS is working to have an additional 225 samples uploaded to the system. NCATS is also setting up agreements (RCAs) with veterinary clinical research teams across the country to assist with the processing and interpreting of genomic sequences and proteomics data, with the understanding that the data will then be uploaded to the ICDC. Another NCATS program, the Clinical and Translational Science Award (CTSA) One Health Alliance (COHA), is engaging veterinary groups to bring them into the human health community. Updates to the COHA website may include portals or links to connect COHA users to the CRDC and the ICDC. NCI hopes to attract more partners in the coming year to continue to build the ICDC repository.

**Overview of the ICDC**

The ICDC is a component of the NCI CRDC that provides a place to share canine cancer data and to connect investigators, researchers, and owners of canine patients. The ICDC is currently in the first year of the 2-year prototype phase of the project, during which the initial canine commons will be built on a robust architecture so that the subsequent production phase can focus on bringing in data and adding new features rather than building the foundation or infrastructure.

Both the CRDC and the ICDC are cloud-based systems. The ICDC is being populated with clinical trials data and operates using a Gen3 technical stack, in which a series of databases (e.g., canine, proteomics, biomarkers, imaging) supports the larger, overarching data commons. With this approach, data are pooled to allow for access across datasets and databases in a reproducible and secure manner. The cloud-based system allows users to access large genomic datasets without needing to download data.

The NCI CRDC includes a set of cloud-based resources, including the Broad FireCloud, the Institute for Systems Biology Cancer Genomics Cloud (ISB-CGC), and the Seven Bridges Genomics (SBG) CGC. These are workspaces for researchers to bring their own data and tools, access pipelines to the repositories, and save and share their data and analyses. The CRDC also includes several tool repositories that can be accessed directly via the cloud resources. The CRDC offers an open-access environment in which users are not limited to individual components of the system, such as the ICDC. Rather, users can access all features, components, and data under the CRDC umbrella. Canine data can therefore be analyzed across human and other species datasets.

Authentication and authorization is achieved through multiple secure APIs that allow researchers to build their own visualization tools. Web interfaces, data submission, and tool deployment will be developed and integrated into the CRDC. A broad range of types of users is expected for both the CRDC and the ICDC.

Three data commons—canine, genomic, and proteomic—currently comprise the components of what is referred to as the harmonized common/metadata model (CRDC-H). Considerable technology goes into building these datasets. In addition to the data, models, digital IDs/metadata services, and APIs are all needed to build the node portal for each of the repositories. The

CRDC-H model will enable users to ask questions across all of the data commons nodes, while the cancer data aggregator (CDA) is the tool that end users will use to ask those questions.

Organizationally, the ICDC has three main components: data, system/infrastructure, and the Steering Committee. The objectives of the data component are to provide a "help desk" and to harmonize and upload data to the commons to build a working ICDC prototype under the infrastructure component. In addition to providing data, the key contributions of the SC members will be to review and provide feedback about use cases, identify community and user features, and review the system as the ICDC prototype is developed. Work on the prototype will set the foundation for the fully operational ICDC, including development of clear guidance for use of the system.

The ICDC prototype is being seeded with trial-based data from NCATS studies and a small-molecule trial from NCI's COP. ICDC staff are working with investigators on those studies to build the necessary tools to extract, transform, and load data from one format (e.g., as submitted as tabulated data or data in PDF files) to another (e.g., a unified output file). ICDC staff will continue to work with researchers to identify the process needed to transform data from the structure used by research teams to a format that matches what is in the system.

The system infrastructure follows a clinical data-based model, which differs from the model used for the GDC. The user interface for the ICDC is under development and will be configured and customized to the ICDC. Users will be able to access information on the homepage, with full log-in access via the main webpage. In contrast with localized data commons, the ICDC is designed to allow for updating of the data model and for users to access individual data subsets and models as well as the overall commons. This approach differs from systems that use a "one-size-fits-all" model.

Questions and discussion

Additional information was requested on how harmonization of canine and human clinical trials data will be done, given how large a task this will be. This aspect of the ICDC ventures into new territory, and the plan for harmonizing data from different sources and species is not yet fully defined. NCI/FNLCR will be looking to the SC to help guide and clarify decisions regarding infrastructure, pipelines, and harmonization, with a focus on data from canine clinical studies. The team will also look to successes and failures from other similar projects—such as Seven Bridges, the GDC, and the Monarch Initiative—as models for how to bring data, teams, and initiatives together under the ICDC. Another resource is the Imaging Data Commons, which has released an NCI-sponsored RFP through FNLCR.

It was noted that cross-pollination and interactions (vs. siloing of data and analyses) are built into the design of the ICDC, with the ultimate aim of facilitating the sharing of data as widely and quickly as possible across teams. The design also presumes some level of iteration and that no one model fits all. Various factors, such as data submission and data querying, will also be taken into account in building the prototype.

Another question focused on whether data from clinical trials will be accepted "as is" or if data will need to be formatted in advance to a uniform standard. It was noted that during the prototype

phase, the team will work with groups that already have data for the commons; in these cases, data will be accepted as submitted and staff will work on extracting and transforming as described above to identify the best tools for this purpose. During the subsequent production phase, the plan is to provide investigators with tools and guidance to facilitate standard formatting of data prior to uploading to the ICDC. In the interim and going forward, the goal is to provide investigators on new projects with guidance on how to format data for the commons to smooth the transition of the data to the ICDC.

**Responses to Questionnaire**

Before the current meeting, Committee members outside NIH were invited to complete a brief questionnaire about the ICDC. Internal members will also be asked to complete the questionnaire. Full results will be shared when available. The compiled responses to date show the following:

- The primary types of data that the ICDC should focus on include DNA and RNA sequences, conserved mutations, and protein sequences. Collection of other types of data did not receive as strong a response.
- Respondents have data they would like to submit to the ICDC.
- The project should track publications as a measure of success.

Responses from both internal and external SC members will help toward the goal of promoting and enabling community access to the commons.

Open discussion

The CRDC includes three parallel data commons: the canine, proteomics, and genomics commons. In response to a question about plans to add data for other animals, with the aim of replicating findings in dogs in other animal models, it was noted that there currently are no plans to expand the ICDC to include other species. At this time, and for the foreseeable future, the focus is on building a canine database to answer the original question of whether pet dogs with spontaneous cancers can serve as models of human disease to assess new immunotherapeutic agents. That research question is the driver for funding of the current effort. Other species and models might be considered in the future, but funding would need to be secured for any expansion.

While there are many opportunities to submit data to repositories, a foundational principle for the current project should be to generate reproducible results that are then shared with the larger community via publication. Therefore, the goals of this project should include not only adding data but also providing opportunities for broad access and analysis across the commons.

Consistent with the intent of the team building the commons and the charge to the SC to define this initiative's measures of success, Committee members suggested tracking usage and outcomes, including short-term measures such as the number of unique hits and the number of users who upload and download data. Such measures will be key to driving investigators to the datasets and, in turn, demonstrating the success of the project. Through tracking usage, the contributions of investigators and the informatics staff can also be acknowledged.

Another model that can be applied involves assigning each unique collection a Digital Object Identifier (DOI) to allow tracking of the data as a publication. This approach is used for The Cancer Imaging Archive, which also has been registered with the Library of Congress as a serial publication (ISSN number 2474-4638), and assigns DOIs for each unique collection (note that it is not required to be a registered publication to assign DOI's). This allows researchers who contribute data to cite the contributed data in TCIA with the corresponding DOI(s) as part of their publication record, as in this example: Schmainda KM, Prah M (2018). Data from Brain-Tumor-Progression. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2018.15quzvnb. Analyses or annotated images can also be provided with a unique DOI. Those who are interested in additional information can then access the data or analysis via the DOI.

Measures that show a return on investment should help drive people to the site and analyze the data in the ICDC datasets.

**Conflict of Interest/Confidentiality Agreement/Honorarium**

Members will need to disclose whether they have any conflicts of interest, which is done by completing a COI form. Dr. Parchment will provide SC members with the required document.

The Committee discussed whether a non-disclosure agreement (NDA) is needed, given that the ICDC is in the public realm. SC members are not currently under a confidentiality agreement. Exceptions that could require an NDA might include discussion and review of RFAs and other funding mechanisms that have not yet been released or announced. Committee members were asked to provide feedback on this issue to Dr. Parchment.

Committee members were reminded to complete and submit their paperwork so that honorarium for their participation and meeting attendance can be processed.

**Future Meetings/Action Items**

Future SC meetings

Convening midweek and midday seems to be best for Committee members. The next meeting has not been scheduled yet but will likely be held in 4 to 6 weeks. Committee members will be polled to identify the best date and time for the next meeting.

Topics for discussion at future meetings should be sent to Dr. Parchment or Mr. Beyers. Committee members were also asked to consider the types of questions they would like to ask the database for the team to discuss over the coming months.

Action items

- Dr. Parchment will distribute the required COI form to SC members.
- External SC members were reminded to forward paperwork for their honorarium to Ms. Lydard.
- Feedback on NDAs for SC members should be forwarded to Dr. Parchment.
- Agenda items for future meetings should be forwarded to Dr. Parchment or Mr. Beyers.

- Dr. Parchment will contact SC members once the next ICDC-SC meeting date and time are finalized.

**Adjournment**

The meeting was adjourned at 12:59 p.m.