



**Q&A and Panel Discussion**  
*PCDC All-Commons Webinar*  
*June 3, 2020*

## Q&A

**Can clinical fellows and graduate students who are not part of COG or a consortium get access to the data commons?**

Yes, access is available to all researchers regardless of affiliation. Anyone can register an account to access the [INSTRuCT](#), [INRG](#), and [MaGIC](#) consortium cohort discovery tools and explore aggregate data numbers. Access to line-level data requires submitting a data request, which must be approved by the consortium Executive Committee before data is released.

**Are there opportunities for trainees (PHO fellows) to be involved even if their institution is not yet part of the commons or a consortium?**

Yes, the PCDC welcomes trainees to join one or more of our work groups (Data Model, Governance, Data Analytics, and Engagement) or to engage with leaders of their particular disease group of interest to identify opportunities for participation, which will vary depending on the progress to date of each disease group. Please contact Monica Palese ([mpalese@uchicago.edu](mailto:mpalese@uchicago.edu)) and Kat Blumhardt ([kblumhardt@uchicago.edu](mailto:kblumhardt@uchicago.edu)) to discuss current opportunities and/or disease group progress.

**Is there specific work to integrate cancer registry and EHR data into the PCDC? If so, what quality assurance and harmonization measures are taken for the data dictionary?**

The PCDC is looking to link any data types that might be useful to researchers. EHR and registry data are particularly important, since they are usually of high quality.

Registry data, often from cancer registries, are manually curated. Quality assurance is very important to the PCDC, so each data source is vetted for policies and procedures that help ensure high quality. In addition, the PCDC runs a series of QA/QC scripts to confirm internal consistency of the data. Our team works closely with data managers and stakeholders from all groups to help with the QC processes.

## **How are "orphaned" rare disease groups capable of getting into a consortium?**

There are three ways in which the PCDC currently incorporates disease groups:

1. The creation of a disease group consortium, managed by the PCDC, comprised of data contributors (e.g., neuroblastoma/INRG, soft tissue sarcoma/INSTRuCT).
2. Storage of data from an external disease group consortium headquartered at another institution (e.g., germ cell tumors/MaGIC).
3. Linkage to data housed by an external disease group consortium (e.g., IDIPGR).

The biggest issue with small disease groups is a lack of funding for commons development. This mostly impacts the data dictionary development. Thus far, the PCDC has been able to help these small groups find funding for data dictionary development and balloting. As more infrastructure is developed, the costs for integration of new commons will decrease considerably.

## **Is governance only applicable prior to release, or is the vision that data distribution will always be governed by individual data contributors?**

Data distribution will always be governed by an executive committee made up of the data contributors. The PCDC is committed to making sure that data contributors have control over the release of their data. This occurs through several governance mechanisms. When cooperative groups (e.g., COG) contribute data to a data commons consortium (e.g., INRG, INSTRuCT), they can put a representative on the consortium's executive committee and thus have oversight for how the data are used for projects. The executive committee governs the use of data as well as quality control for contributed information. The executive committee also adopts a publication policy that is consistent with the needs and values of the various cooperative group representatives. Governance is a key element of the success of the data commons and remains one of the top priorities.

## **If an investigator is interested in asking a question that crosses multiple disease types (e.g., survival data for a genomic finding found in both liquid and solid tumors), would they need to obtain permission from each relevant consortium before performing that query?**

Yes, but that process will be streamlined. The governance plan currently being developed and refined will include these project requests coming through the PCDC Executive Committee but requiring approval of the individual disease group executive committees. While this may appear onerous, it is the only way to ensure that each disease group retains their autonomy in deciding how their data are used for research. Thus far, the disease consortia have supported this vision.

**Once a project is done, is there required review by the consortia prior to publication?**

We are currently discussing and documenting the governance policies for publication. Each consortium is developing their own publication policy, which will govern how data can be used for publication and how attribution is assigned. Some consortia are requiring review of the paper prior to publication, while another approach is requiring the participation of a consortium-approved statistician in the analysis.

**Are probabilistic linkages permitted on European data, since identifiable data is not permitted?**

GDPR does not allow an identifier to be associated with the data. This makes linkages very difficult, and they can currently only be accomplished using other disambiguation techniques. It is possible that the University of Chicago could host data with identifiers, but we would then need to adhere to GDPR “right to forget” standards. European consortia are starting to use the commercial service EuPID to assign identifiers and to associate related data sets.

**In the current state, when a researcher requests data, where does the funding for the effort to pull the data come from?**

PCDC funding currently relies on private donations and grants, which we are actively seeking. Thus far, all data pulls have been performed without cost to the researchers, and the PCDC hopes to maintain this model for the next few years. Ultimately, a sustainability model will likely include some cost to researchers for data. Other support may come from cooperative groups.

**For a long-term effects patient study (PCORI-funded), who would house the commons for 10+ years?**

The long-term plan is for the PCDC to continue to host data. The infrastructure costs for the commons are a small fraction of the overall cost. The PCDC plans to continue hosting clinical data in a cloud-based infrastructure. Our adherence to common standards means that the data could easily be moved to another platform/group if needed. Ultimately, the PCDC should be sustainable with a combination of philanthropic, cooperative group, commercial, and governmental funds. The PCDC team has experience with the PCORI initiative, specifically the CAPRiCORN (Chicago PCORI) project, and where possible, the PCDC is leveraging the standards work for PCORI.

## Panel Discussion

**Sam Volchenbom, MD, PhD:** *(for Sue Cohn)* Talk about the challenges you and the INRG group had to overcome to get the data commons off the ground. How relevant is this to our work now?

**Sue Cohn, MD:** This work was started more than fifteen years ago in 2005. Neuroblastoma already had a culture of international collaboration, because it is such a rare tumor, that was rare in other disease groups at the time. For this reason, neuroblastoma researchers have been collaborating for many decades to try to come up with common terms. However, the existence of such a large cohort of data presented a new challenge for international collaborators, which had been relatively simple prior to combining all data.

Over the last few years, the rules have changed substantially, with additional protections like HIPAA and EU privacy laws. These protections may pose additional challenges if the neuroblastoma commons were trying to get off the ground today.

Another challenge that the group has faced is in trying to make a living cohort of data. There have been efforts to keep the commons updated with new diagnoses and outcomes, and we have been successful in collaborations with COG and the statistical office. It hasn't been quite as simple to make these changes with our European counterparts, however, due to differences in rules and regulations. We have been able to enter new patients from European studies, but there have been difficulties in updating existing outcome data. There have also been challenges in linking to other data sets and, therefore, we have fewer data linked to genomic, pathologic, or radiographic datasets.

Overall, this resource has enabled new discoveries in neuroblastoma research that could not have been made with smaller cohorts. Even in the initial phases of the project when INRG still lived on Excel datasheets, there were important papers published with new biomarkers and new ways of thinking about stage, among other things. I am hopeful that as more visualization tools are integrated, there will be increased knowledge-sharing and discovery and as we increase linkages to genomic data, the data will become richer and more useful.

**Sam:** *(for Julie Guillot)* Comment on how the PCDC can start to leverage parents and parent groups for advocacy and help in building commons and finding funding sources.

**Julie Guillot:** I was recently at a meeting where one of presenters said that it's much harder to cure cancer than to send someone to the moon. I believe that we can cure childhood cancer, but there are challenges facing the cancer research community that NASA (for example) does not have to deal with. NASA is fully funded and has a mandate. They are able to leverage the skills of the best engineers, physicists, data scientists, and

mathematicians. For the cancer research community, there is a challenge in identifying parents and organizations to help with both funding and engagement opportunities.

One thing the AML team has done for me is to spend time and resources to educate me and include me in discussions. This attitude of inclusion has helped me to learn and get energized. I think it is important that the research community includes and embeds us (parents and advocates). When I am included in discussions with the Leukemia and Lymphoma Society (LLS), I can bring a different set of perspectives and opportunities. Parents and advocates are sources of fundraising, but we are also great problem solvers who come from a variety of backgrounds. We are often able to come up with unique ideas that would not otherwise be considered. Connect, identify, educate, and invest in us. Include us. Embed us. Don't discount us.

To parents, I would implore you not to wait to be invited. Be proactive, reach out. If the PCDC includes something you have a background in (like tech, data, or contacts), get involved with Sam and his team and see if you can help. Bring your creativity and crazy ideas. I think there is a great opportunity for a formal PCDC advocacy group. This group could look at funding sources and attempt to increase participation and partnerships.

I think this whole initiative is very exciting and it will be able to drive huge discovery.

**Sam:** *(for Katie Janeway)* You have done a lot of work with EHR (electronic health records) to leverage data. Briefly mention how EHR will help with pediatric cancer data research.

**Katie Janeway, MD, MMSc:** We've been conscientious about principles in data sharing. In particular, we have been trying to work in concert with other groups who are also developing data standards. This includes close communication with the PCDC and each disease group.

There are some amazing benefits of working with EHR data. When using data from an EHR system, you have access to data that are comprehensive. We are interested in precision oncology and genomics. I work at an institution where every patient has had their cancer sequenced and we wanted to annotate this information with clinical data. We were able to do this using the EHR because we can go into the entirety of each patient's medical course. The challenge is to do it in a way that has high quality and reproducibility.

We want our data to be harmonized with registry data and clinical trial data, but this is a work in progress. We currently have ongoing discussions with the PCDC, Dr. Pembrath at the National Childhood Cancer Registry, the National Childhood Cancer Data Initiative.

There are several institutions contributing pediatric sequencing data, including Dana-Farber, MSK, CHOP, and UCSF. I am happy to talk more with anyone who is interested!

**Sam:** *(for Mignon Loh)* Talk about how collecting data will help the ALL community, since ALL is the most common type of pediatric cancer. What will the benefits be?

**Mignon Loh, MD:** A comprehensive approach to integrate clinical data with trial data and genomics is one of the strengths of genomic data. The PCDC provides an opportunity for maximal statistical power and the ability to dive deeply into subsets and outcomes.

Beyond that, I am also hoping to enjoy the benefits of working and collaborating with international partners. We are already conducting more international studies and the work we are doing on the commons will help to inform how we conduct these international projects and provide a common platform to make international data more nimble.

**Sam:** *(for Doug Hawkins)* As far as future development of clinical trials, how do you see data commons helping Children's Oncology Group (COG) with development of ongoing trials and data collection?

**Doug Hawkins, MD:** Many people say COG is great but it takes too long, it's too slow. The data commons gives us an opportunity to standardize how we collect data so that we don't need to reinvent the wheel every time we conduct a study. This will allow us to collect common elements the same way and speed up the timeline from activation to time of study. It will also make the analysis of the study easier and allow for data cleaning as the study is conducted, leading to a manuscript more quickly.

Even within COG we have siloed data that uses different standards. If we can establish a dataset that can be used even within COG to speed up our processes, it will lend itself to the ability to collaborate with international partners. In the end, we will be able to speed up the exchange of data and save lots of time.