JAMIA Open, 2024, **7(1)**, ooae004
https://doi.org/10.1093/jamiaopen/ooae004
Case Report

Case Report

An open-source platform for pediatric cancer data exploration: a report from Data for the Common Good

Kirk D. Wyatt, MD[1,2], Luca Graglia, MS, MBA[2], Brian Furner, MS[2], Bobae Kang, MA[2], Michael Fitzsimons, PhD[3], Robert L. Grossman, PhD[3], Samuel L. Volchenboum, MD, PhD[2,4,*]

[1]Department of Pediatric Hematology/Oncology, Roger Maris Cancer Center, Sanford Health, Fargo, ND 58102, United States, [2]Data for the Common Good, University of Chicago, Chicago, IL 60637, United States, [3]Center for Translational Data Science, University of Chicago, Chicago, IL 60637, United States, [4]Department of Pediatrics, University of Chicago, Chicago, IL 60637, United States

*Corresponding author: Samuel L. Volchenboum, MD, PhD, Department of Pediatrics, University of Chicago, 900 E. 57th Street, Chicago, IL 60637, United States (slv@uchicago.edu)

Abstract

Objective: The Pediatric Cancer Data Commons (PCDC)—a project of Data for the Common Good—houses clinical pediatric oncology data and utilizes the open-source Gen3 platform. To meet the needs of end users, the PCDC development team expanded the out-of-box functionality and developed additional custom features that should be useful to any group developing similar data commons.

Materials and Methods: Modifications of the PCDC data portal software were implemented to facilitate desired functionality.

Results: Newly developed functionality includes updates to authorization methods, expansion of filtering capabilities, and addition of data analysis functions.

Discussion: We describe the process by which custom functionalities were developed. Features are open source and available to be implemented and adapted to suit needs of data portals that utilize the Gen3 platform.

Conclusion: Data portals are indispensable tools for facilitating data sharing. Open-source infrastructure facilitates a modular and collaborative approach for meeting needs of end users and stakeholders.

Lay Summary

Data commons provide information technology infrastructure to house data and facilitate sharing and analysis of data within a unified ecosystem. The Pediatric Cancer Data Commons facilitates sharing of pediatric cancer data and utilizes the Gen3 data platform. To meet the needs of end users, the Gen3 platform was customized, including updates to authorization methods, expanding filtering features, development of a Kaplan-Meier survival analysis tool, and linkage to external data commons. These customizations are available as open-source software for other data commons using the Gen3 platform to utilize.

Key words: data commons; pediatric oncology; health information management (MeSH); data visualization (MeSH).

Objective

To describe the development of custom features for a data commons housing clinical pediatric cancer data.

Background and significance

Data commons provide information technology infrastructure to house data and facilitate sharing and analysis of data within a unified ecosystem.[1,2] The Pediatric Cancer Data Commons (PCDC; pedscommons.org)—a project of Data for the Common Good (dataforthecommongood.org)—includes clinical data on children with cancer.[3] As of this writing, data on nearly 40,000 patients are housed in the PCDC. Data within the PCDC are largely sourced from prospective clinical trials but are anticipated to increasingly include registry data. Data fields include demographics, diagnosis, prognostic features, molecular alterations, and outcomes (eg, relapse, death). The PCDC utilizes an open-source data commons platform (Gen3; gen3.org) and developed a number of customizations to suit the needs of stakeholders. Herein, we describe the development of the PCDC Data Portal and features developed to facilitate data exploration by end users.

PCDC architecture and data portal functionality

The Gen3 data platform includes core functionality needed to host a data commons ecosystem and provides flexibility to choose only the components needed, along with support for APIs to build new functionality on top of existing components. Key components of the Gen3 platform are summarized in Supplementary Appendix 1. While Gen3 did not fully suit the needs of the PCDC "out-of-box," it was the preferred platform, as it served core functions using open-source software. By adopting a well-established, open-source platform to host a data commons, the PCDC technical team could focus on customizing the platform to meet the needs of end users.

Received: May 24, 2023; **Revised:** October 30, 2023; **Editorial Decision:** January 5, 2024; **Accepted:** January 8, 2024

Table 1. Summary of modifications.

Description	Problem	Solution
Authorization		
Granular authorization	Authorization extended only to high-level node (versioned dataset)	Extend authorization to lower-level node (patient)
Data request	Lack of ability to track data request status	Develop service to associate users with data requests with individualized permissions
Filtering		
Saved filter sets	Filtering selections could not be saved for future reference	Leverage state service to persist filter selections for future retrieval
Filter selection within variables	Multiple value selections within a variable can only be combined with OR logic	Add toggle to implement option of NOT IN logic to combine values
Filter set workspace	Need for management of saved filter sets and complex composition of filter sets	Develop user interface and back-end functionality to manage and compose filter sets
Sharing saved filter sets	Inability to share filtersets across users	Develop service to generate and ingest unique tokens, which are associated with filter selections
Data analysis		
Kaplan-Meier survival analysis tool	Lack of on-platform analytics tools	Develop back-end microservice to develop Cartesian points and front-end service to draw survival curve
Other functionality		
Custom ETL service	Existing ETL service had limited support for nested objects	Develop custom ETL engine using on-demand cluster
Linkage to external data commons	Inability to link to other datasets	Connected identifiers across multiple platforms

As the PCDC Data Portal was being prepared for public launch—and after the initial launch—a number of desired customizations were identified and implemented. We review these key customizations below (Table 1).

Modifications to gen3

Authorization

Granular authorization

The default implementation of Gen3 allows users access to data on a "project" level. Within the PCDC, "project" refers to a specific data release version of the data available on Gen3. For this reason, a user's access to data/subjects within a versioned dataset would be all-or-none. The PCDC maintains different versions of the data to be able to recreate cohorts at different time points. In our implementation, users are generally approved to access data on a specific cohort of subjects instead of all subjects in the versioned dataset. To facilitate providing access to only specific patients, we extended user authorization so that it could be applied in the graph database down to the individual patient-level node. This required updates to most services that work with data in the environment and was facilitated by Gen3's policy engine (Arborist).

Data request

The PCDC data portal must maintain controlled access to line-level (ie, individual patient-level) data and support a system for data requests to the individual consortia that make up the PCDC. Given the complexities of supporting granular access control and individual consortium processes, it was important to develop technical solutions that support scalability and sustainability through automation. To automate our existing data request process, we developed a service (Amanuensis; github.com/chicagopcdc/amanuensis) to handle user state data. Amanuensis was designed to support flexible user permissions, enabling one or more users to be associated with a data request with individualized permissions (eg, to allow specific users to only see the status of a request [eg, PCDC administrative staff] or to also see the data once they are available [eg, data requestor]). If a project request is approved, the service generates a Portable Format for Bioinformatics file containing the data the user has been approved to access.

Filtering

Saved filter sets

The Gen3 interface allows users to select cohorts of interest by applying faceted filters. One limitation of native Gen3 filtering functionality is that filtering selections cannot be saved for future reference and have to be individually re-applied each time a user wishes to review a cohort of interest.

To allow filter sets to be saved, we extended Amanuensis, the user state data service. To support saving of filter sets, Amanuensis persists the state of the exploration page (ie, selected filters) in a JavaScript Object Notation blob, to which the user is able to associate a name and a description. Other services can then retrieve this saved state for state-related actions, including loading a saved filter set, initiating a data request, and generating a survival curve.

Filter selection within variables using NOT in and ALL operators

By default, selecting a value within a filter set will include subjects where the variable value is equal to the value selected. When multiple values are selected, these are combined using the OR operator (eg, selecting values of "Alveolar rhabdomyosarcoma" and "Alveolar soft-part sarcoma" under Histology will yield patients with either of these histologies; Figure 1). To simplify filter selection, an "exclude" toggle button was added to each faceted filter variable. When selected, the NOT IN operator is appended to the selected values (instead of the default OR operator) to allow users to select which filter values to exclude from the result.

Filter set workspace

A more modular and user-extensible solution for managing filter sets—termed "Filter Set Workspace"—was developed.

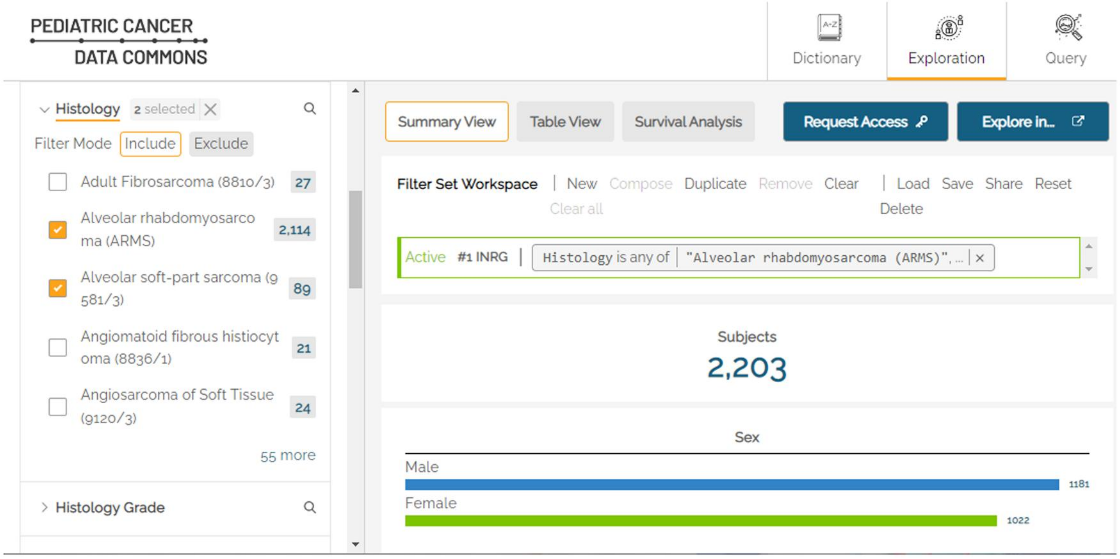

PEDIATRIC CANCER
DATA COMMONS
Dictionary
Exploration
Query
Histology 2 selected
Filter Mode Include Exclude
□ Adult Fibrosarcoma (8810/3) 27
☑ Alveolar rhabdomyosarcoma (ARMS) 2,114
☑ Alveolar soft-part sarcoma (9581/3) 89
□ Angiomatoid fibrous histiocytoma (8836/1) 21
□ Angiosarcoma of Soft Tissue (9120/3) 24
55 more
Histology Grade
Summary View
Table View
Survival Analysis
Request Access
Explore in...
Filter Set Workspace | New Compose Duplicate Remove Clear | Load Save Share Reset
Clear all
Delete
Active #1 INRG | Histology is any of "Alveolar rhabdomyosarcoma (ARMS)", ... ×
Subjects
2,203
Sex
Male
1181
Female
1022

Figure 1. When "Include" filter mode is selected, values are combined using OR logic. When "Exclude" filter mode is selected, the NOT IN operator is applied.

Filter Set Workspace built upon saved filter sets by introducing the ability to quickly toggle between and manage multiple filter sets at once. Filter Set Workspace also introduced the ability to develop "Composed" filter sets that represent combinations of created filter sets using AND/OR logic. The resulting filter sets of such composing operations can then be further composed to generate filter sets of even greater complexity. For example, users could select for patients who have neuroblastoma AND have either chromosome 1*p* OR 11*q* deletion (Figure 2).

Using the compose feature, a user can save multiple filters and then compose them (ie, connect them with AND/OR) to create more specific filters. When two or more filters are composed, the application generates a new root-level component that connects the selected filters as child components.

Sharing saved filter sets

Another feature related to the Filter Set Workspace is the ability to share saved filter sets with other users. When a user selects the "Share" functionality, they are presented with a unique token (generated by a backend service) and the ability to copy the token to the clipboard which can then be shared with another user. The recipient of the token can paste the token into a box on the "Load" filter screen to load a copy of the shared filter set.

Data analysis

Kaplan-Meier survival analysis tool

Kaplan-Meier survival analysis is often used as the primary analysis for clinical trials in pediatric oncology to determine treatment efficacy. Development of a front-end data visualization tool for performing Kaplan-Meier survival analysis was identified by clinicians and researchers as a key desired functionality. The tool (Figure 3) allows users to plot survival for any cohort represented in a saved filter set.

The front-end component is responsible for two tasks:

1) Collecting the input/parameters from the user and forwarding them to the backend (ie, what cohort(s) of patients to analyze, what time frame to show, overall vs. event-free survival)
2) Drawing the curve using the result received from the backend

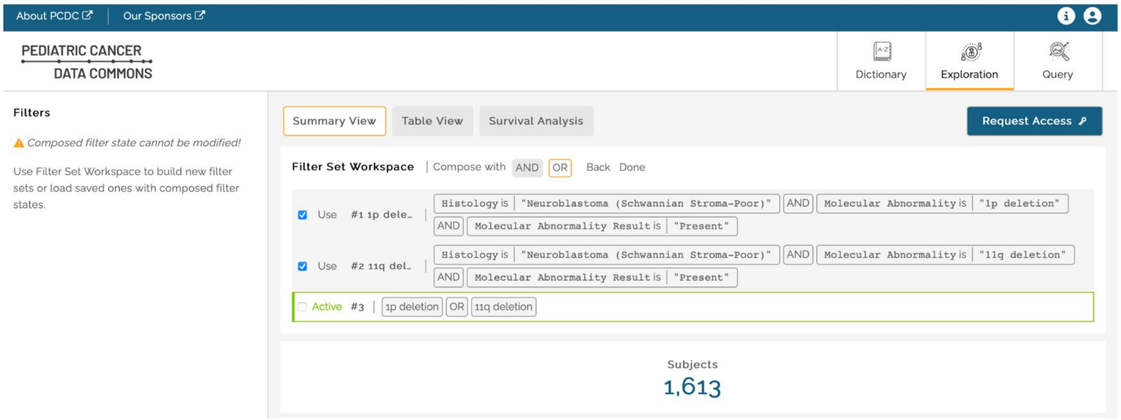
The generation of data points for the curve is performed by a backend microservice that interacts with the data source. This service returns to the front end only the final generated Cartesian points to interpolate and draw the curve. In this way, line-level data are not exposed to the user. This service is a Python Flask app and uses the Lifelines survival analysis package.[4]

As the functionality of the Kaplan-Meier survival analysis tool was shared with key stakeholders, concerns were raised regarding the potential that the tool could potentially be misused or abused by users (eg, "*p*-hacking"). In light of these concerns, a number of safeguards were put in place. Because some consortia were not comfortable with data being explored using the Kaplan-Meier survival analysis tool—even with implemented safeguards—functionality was requested to allow survival analysis only for data represented by consortia that opt-in. Consortia also expressed a desire to allow filtering by variables, such as data contributor, research study, and study arm to facilitate sample size estimations but to disallow generation of survival curves that include granular data selected according to these variables. The primary rationale for preventing survival analysis when filtering by these variables was to avoid the possibility that users could perform short-sighted analyses comparing countries or cooperative groups that could undermine collaborative data-sharing efforts. To facilitate this functionality, "allowed consortia" and "disallowed variables" lists were created for the survival analysis tool. Filter sets that include consortia not on the "allowed consortia" list or which filter by variables on the "disallowed variables" list appear grayed out and are not selectable in the filter set selection drop-down box when generating survival curves.

Other functionality

Custom extract, transform, load service

The default extract, transform, load (ETL) service (Tube; github.com/uc-cdis/tube) is elastic and suitable for many use



Filters

⚠ Composed filter state cannot be modified!

Use Filter Set Workspace to build new filter sets or load saved ones with composed filter states.

Filter Set Workspace | Compose with AND OR

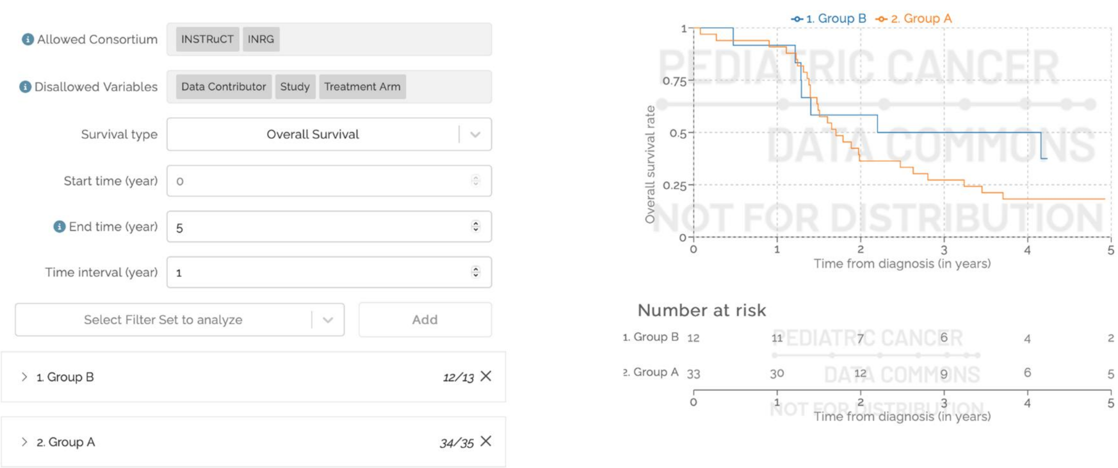
☑ Use #1 1p dele… | Histology is "Neuroblastoma (Schwannian Stroma-Poor)" AND Molecular Abnormality is "1p deletion" AND Molecular Abnormality Result is "Present"

☑ Use #2 11q del… | Histology is "Neuroblastoma (Schwannian Stroma-Poor)" AND Molecular Abnormality is "11q deletion" AND Molecular Abnormality Result is "Present"

□ Active #3 | 1p deletion OR 11q deletion

Subjects 1,613

Figure 2. Filter Set Workspace with composed filter set.


Allowed Consortium: INSTRuCT INRG

Disallowed Variables: Data Contributor Study Treatment Arm

Survival type: Overall Survival

Start time (year): 0

End time (year): 5

Time interval (year): 1

Select Filter Set to analyze | Add

1. Group B 12/13

2. Group A 34/35

Figure 3. Current Kaplan-Meier Survival Analysis Tool, with watermark included to discourage unauthorized distribution of survival curves.

cases, but support for nested objects—which was an important use case for the PCDC—was limited, and the way Tube was designed made it challenging to adapt to our use case. For instance, the object shown below could not be generated:

```
{
    "Subject_id": "1",
    ...
     "Tumor_characteristics": {
           "Tumor_assessment": "",
           "Disease_phase": "initial diagnosis",
           ...
    },
    ...
}
```

The complexity lies not only in nesting the list of tumor_assessment within the subject object, but also flattening the disease phase (timing reference) object into the nested object in order to have the time of the event and the event in the same object to simplify the Elasticsearch query translation from the application state.

To solve this issue, we developed a custom ETL engine that utilizes an on-demand Amazon Web Services EMR cluster.

Linkage to external data commons

We have linked clinical data in the PCDC to the corresponding genomic data in the Gabriella Miller Kids First Data Resource and the National Cancer Institute's Genomic Data Commons portal using the Children's Oncology Group Universal Specimen Identifier. Once a user identifies a cohort of interest, the PCDC Data Portal displays how many of the subjects are represented in external data commons, and users are able to click the "Explore In…" button to link out to a display of the overlapping cohort in the external data commons.

Discussion

Benefits and future direction

The end result of the customizations to the Gen3 platform is a data portal that serves the needs of our user base. As the PCDC grows and its user base expands, we expect goals for future functionality to evolve. Our team is currently performing formal usability testing to identify usability challenges and develop user interface improvements to address them.

Another key project in process is automation of the data request process. As the current data access request process is handled on a consortium-by-consortium basis with varying submission formats (eg, fillable Word document or webbased form), we aim to develop a singular data request process and more deeply integrate a centralized request system within the data portal, with the aim to keep data-related tasks within the same ecosystem.

We also plan to build upon filter set sharing functionality by creating bidirectionally synchronized "active" filters that can be updated by any authorized user. We also plan to allow for shareable pre-selected "public" filter sets of common interest to be available to any user in a "filter set library." We also anticipate further increasing the granularity with which cohorts can be selected by adding the ALL operator to the Compose feature of Filter Set Workspace, as described earlier.

Conclusion

As of October 2023, the PCDC includes data on nearly 40 000 patients. We were able to leverage the built-in capabilities of Gen3 to allow us to allocate development resources on customized functionality. Other groups building similar data commons may leverage these added capabilities, which are open source.

Author contributions

Conceptualization: KW, LG, and SV. Software: BK, LG, and BF. Writing—original draft: KW and LG. Supervision: SV, RG, and MF. Funding acquisition: SV. Writing—review & editing: KW, LG, BF, BK, MF, RG, and SV. All authors contributed substantially to the work, reviewed the manuscript critically, approve of the final version to be published, and agree to be accountable for all aspects of the work.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by St Baldrick's Foundation Research Grant Award 585226.

Conflicts of interest

K.D.W. receives subaward funding relating to work on the PCDC. L.G., B.F., and S.L.V. are employed by University of Chicago and contribute to the PCDC in the course of their employment. M.F. and R.L.G. are employed by University of Chicago and contribute to the Gen3 platform in the course of their employment.

Data availability

Repositories for described functionality are publicly available on GitHub at github.com/chicagopcdc.

References

1. Grossman RL, Heath A, Murphy M, Patterson M, Wells W. A case for data commons: toward data science as a service. *Comput Sci Eng*. 2016;18(5):10-20.
2. Grossman RL. Ten lessons for data sharing with a data commons. *Sci Data*. 2023;10(1):120.
3. Plana A, Furner B, Palese M, et al. Pediatric cancer data commons: federating and democratizing data for childhood cancer research. *JCO Clin Cancer Inform*. 2021;5(5):1034-1043.
4. Davidson-Pilon C. Lifelines: survival analysis in Python. *JOSS*. 2019;4(40):1317.