

The Gabriella Miller Kids First Data Resource for genomic research in pediatric cancer and congenital anomalies

David M. Higgins,¹ Christopher Blackden,¹ Miguel Brown,¹ Christina Diaz,¹ Leslie Duffy,¹ Christopher Friedman,¹ Yiran Guo,¹ Mateusz Koptyra,¹ Qi Li,¹ Alex Lubneuski,¹ Dan Miller,¹ Bobby Moulder,¹ Wendy Payton,¹ Emily Reed,¹ Whitney Rife,¹ Alex Sickler,² Natasha Singh,¹ Amanda Warkow,¹ Eric Wenger,¹ Jessica Wong,¹ Bo Zhang,¹ Chuwei Zhong,¹ Yuankun Zhu,¹ Gaele Altefrohne,³ Denis Beauregard,³ Jeremy Costanza,³ Evans Girard,³ Luc-Frederic Langis,³ Lucas Lemonnier,³ Adrian Paul,³ Céline Pelletier,³ Karine St-Onge,³ Jean-Philippe Thibert,³ Michele Mattioni,⁴ Surya Saha,⁴ Jared Rozowsky,⁴ Milos Trboljevac,⁴ Marko Zecevic,⁴ Jeff Knight,⁵ Gina Kuffel,^{5,6} Radhika Reddy,⁵ Pauline Ribeyre,⁵ Bailey Farrow,⁷ Yelena Cox,⁸ Andrew Ericson,⁹ James Ferreris Vargas,⁹ Sophie Forman,⁹ Sofia Labrecque,⁹ Eric Torstenson,⁸ Kai Yin Ho,⁹ Nicole A. Vasilevsky,¹⁰ Emily Boja,¹¹ Valerie Cotton,¹² Danielle Dae,¹³ Marcia V. Fournier,¹⁴ Jaime M. Guidry Auvil,¹⁵ Andréa C. Harris,¹⁴ Huiqing Li,¹⁶ Lu Wang,¹⁷ Brandi Davis-Dusenbery,⁴ Melissa A. Haendel,¹⁸ Jena Lilly,¹ Phillip B. Storm,^{1,19} Sam Volchenboun,⁷ Robert L. Grossman,⁵ Robert Carroll,⁸ Jack DiGiovanna,⁴ Deanne Taylor,²¹ Vincent Ferretti,^{3,20} Adam Resnick,^{1,*} and Allison P. Heath^{1,*}

Summary

Nine-year-old brain tumor patient Gabriella Miller challenged members of Congress to “stop talking and start doing” when providing federal funding for research into cures for pediatric cancer and congenital anomalies. Though she ultimately lost her life to that cancer, her advocacy efforts resulted in the 2014 Gabriella Miller Kids First Research Act, launching the Gabriella Miller Kids First Pediatric Research Program at the National Institutes of Health (NIH). The overarching goal of the Gabriella Miller Kids First Pediatric Research Program is to help researchers uncover new insights into the biology of childhood cancer and congenital anomalies. Following the signing of the Gabriella Miller Kids First Research Act 2.0 in January 2025, the program has been extended at NIH through 2028 to advance the groundwork laid in the program’s first ten years. The Gabriella Miller Kids First Data Resource Center has since honored her legacy by building a comprehensive data resource for genomic research into pediatric conditions. Data from more than 30,000 participants annotated with demographic and clinical information related to their diagnoses have been released for secondary research and analysis using the center’s web-based platforms. This paper analyzes the outcomes of the initiative and highlights breakthroughs made by the larger research community resulting from the availability of this data resource. We explore the future expansion of the data resource to include new modalities and tools for supporting life-saving research for children like Gabriella Miller.

Introduction

Pediatric cancer and congenital anomalies are the two leading causes of death by disease in children up to age 14.¹ Only 1% of cancer cases in the United States occur in people younger than 20 years old, making pediatric cancer much rarer than cancer in adults.² Pediatric cancers differ from adult cancers at the genetic level, with somatic

mutations playing a smaller role and germline mutations in cancer predisposition genes playing a larger role compared to adult cancers. It is hypothesized that these differences are due to the embryonic origin of many pediatric cancers, with their genetic causes lying in the dysregulation of developmental pathways.³

Congenital anomalies, sometimes called structural birth defects, are more common than pediatric cancer,

¹Center for Data-Driven Discovery in Biomedicine, Children’s Hospital of Philadelphia, Philadelphia, PA, USA; ²Children’s National Hospital, Washington, DC, USA; ³Centre Hospitalier Universitaire Sainte-Justine, Montréal, QC, Canada; ⁴Velsera, Charlestown, MA, USA; ⁵Center for Translational Data Science, University of Chicago, Chicago, IL, USA; ⁶Leidos Biomedical Research, Ballenger Creek, MD, USA; ⁷Department of Pediatrics, University of Chicago, Chicago, IL, USA; ⁸Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; ⁹Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, USA; ¹⁰Critical Path Institute, Tucson, AZ, USA; ¹¹Office of Data Sharing, National Cancer Institute, Rockville, MD, USA; ¹²Office of Data Science and Sharing, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA; ¹³Genomic Epidemiology Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA; ¹⁴Developmental Biology and Congenital Anomalies Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA; ¹⁵Office of Data Sharing, National Cancer Institute, Bethesda, MD, USA; ¹⁶Heart Development and Structural Diseases Branch, National Heart, Lung, and Blood Institute, Bethesda, MD, USA; ¹⁷Office of the Director, National Institute of Dental and Craniofacial Research, Bethesda, MD, USA; ¹⁸Department of Genetics, University of North Carolina School of Medicine, Chapel Hill, NC, USA; ¹⁹Division of Neurosurgery, Children’s Hospital of Philadelphia, Philadelphia, PA, USA; ²⁰Department of Biochemistry and Molecular Medicine, University of Montréal, Montréal, QC, Canada; ²¹Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, Philadelphia, PA, USA

*Correspondence: resnick@chop.edu (A.R.), heathap@chop.edu (A.P.H.)

<https://doi.org/10.1016/j.ajhg.2026.07.010>

© 2026 The Authors. Published by Elsevier Inc. on behalf of American Society of Human Genetics.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

estimated to affect 3% of births in the United States.⁴ The most prevalent congenital anomalies include heart defects such as pulmonary valve atresia and stenosis, trisomy 21, cleft lip with and without cleft palate, and clubfoot.⁵ Congenital anomalies are the leading cause of death in infants less than 1 year old and decrease the quality of life in both the children diagnosed with them and their parents and caretakers.^{1,6,7} The genetic causes of congenital anomalies are as multiform as the conditions themselves, with studies identifying many different genes associated with conditions such as orofacial clefts and congenital diaphragmatic hernia.^{8,9}

Pediatric cancer and congenital anomalies are linked, with several studies indicating that people born with a congenital anomaly are more likely to be diagnosed with cancer as children and adults.^{10–12} Continued research into the identification of the shared genetic pathways of these conditions is essential for the development of therapies for the children who receive these life-changing diagnoses.

Gabriella Miller was one of these children. Nine-year-old Gabriella was diagnosed in 2012 with diffuse intrinsic pontine glioma (DIPG), an aggressive type of pediatric brain tumor with a median overall survival of 11.2 months following diagnosis despite decades of research.^{13,14} Gabriella became an outspoken advocate for fellow pediatric patients with cancer while receiving treatment, starting her own research foundation and challenging the members of the United States Congress to “stop talking and start doing” for pediatric cancer research by increasing federal funding for diseases such as DIPG.

Although Gabriella Miller passed away less than a year after her diagnosis, her advocacy efforts resulted in the signing of the Gabriella Miller Kids First Research Act in 2014, which allocated \$126 million over 10 years for pediatric research. In the decade since, the Gabriella Miller Kids First Data Resource Center (Kids First DRC) has worked to honor Gabriella’s legacy as a collaborative pediatric research effort to understand the genetic causes of and links between childhood cancer and congenital anomalies.

The goal of this perspective is to highlight the successes, challenges, and limitations of building such a data resource. Clinical, phenotypic, or biospecimen information for 38,180 participants has been publicly released as of the end of 2025. Of this information, genomic data are available for 30,854 participants, all hosted on cloud-based platforms to support analysis for authorized researchers. Here, we describe how genomic files are generated and processed from patient-consented samples, are associated with clinical metadata, and are made available via our cloud-based infrastructure. We also describe some of the program’s unique features, place Kids First in context with other large genomics consortia, and present a summary analysis of the peer-reviewed publications using Kids First data. We conclude with a look to our future on how we improve upon what we have built and how we

can connect these basic science discoveries to impacts in the clinic.

Our datasets

The Kids First DRC curates and harmonizes multimodal data (clinical, genomics, and transcriptomics) from selected research studies into a single data resource across pediatric conditions. Each study originates as an independent research project focused on a specific condition or a set of related conditions (Figure 1). Investigators leading these projects with participants who have consented to allow broad research use of their deidentified genomic data can apply to the National Institutes of Health (NIH) for sequencing support through the X01 resource access program.

Genomic and other omics data files are processed using a core set of bioinformatics pipelines. By running each Kids First study through the same bioinformatic pipelines, the output files are rendered functionally equivalent and suitable for combined, cross-study analysis within the program. Kids First’s structure of individual cohorts of research participants harmonized together for combined analysis is similar to other large-scale publicly funded human genomics research data platforms, such as the National Cancer Institute (NCI)’s Genomic Data Commons (GDC)¹⁵ and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) programs.¹⁶ Kids First built upon some of the successes of TARGET and shares similarities with the older program, including the release of multiple output formats from standardized bioinformatic workflows. Kids First has an expanded scope compared to TARGET, including congenital anomaly studies and genomic data from unaffected family members, which can be used for trio-based *de novo* variant identification.

Each genomic file is accompanied by a collection of metadata information about its associated participant(s) and biospecimen(s). Clinical data fields are received from the original investigator and harmonized for cross-study search and cohort development using two ontologies to describe participants’ clinical conditions: the Human Phenotype Ontology (HPO) for phenotypes¹⁷ and the MONDO Disease Ontology for diseases.¹⁸ Each of these ontologies is organized hierarchically, where specific conditions are nested within broader conditions, enabling searching at multiple levels for combining cohorts of patients across studies.

Currently, there are 36 Kids First studies available for research use, representing data from more than 30,000 participants, each of which can be explored on the Kids First Data Resource Portal (see [our integrated platforms](#)). Summary-level information about the participants in the Kids First Portal with either whole-genome sequencing (WGS) or RNA sequencing (RNA-seq) data files is presented in Table 1. The Kids First DRC shares self-identified

GABRIELLA MILLER KIDS FIRST DATA RESOURCE CENTER DISEASE AREAS OF INTEREST:

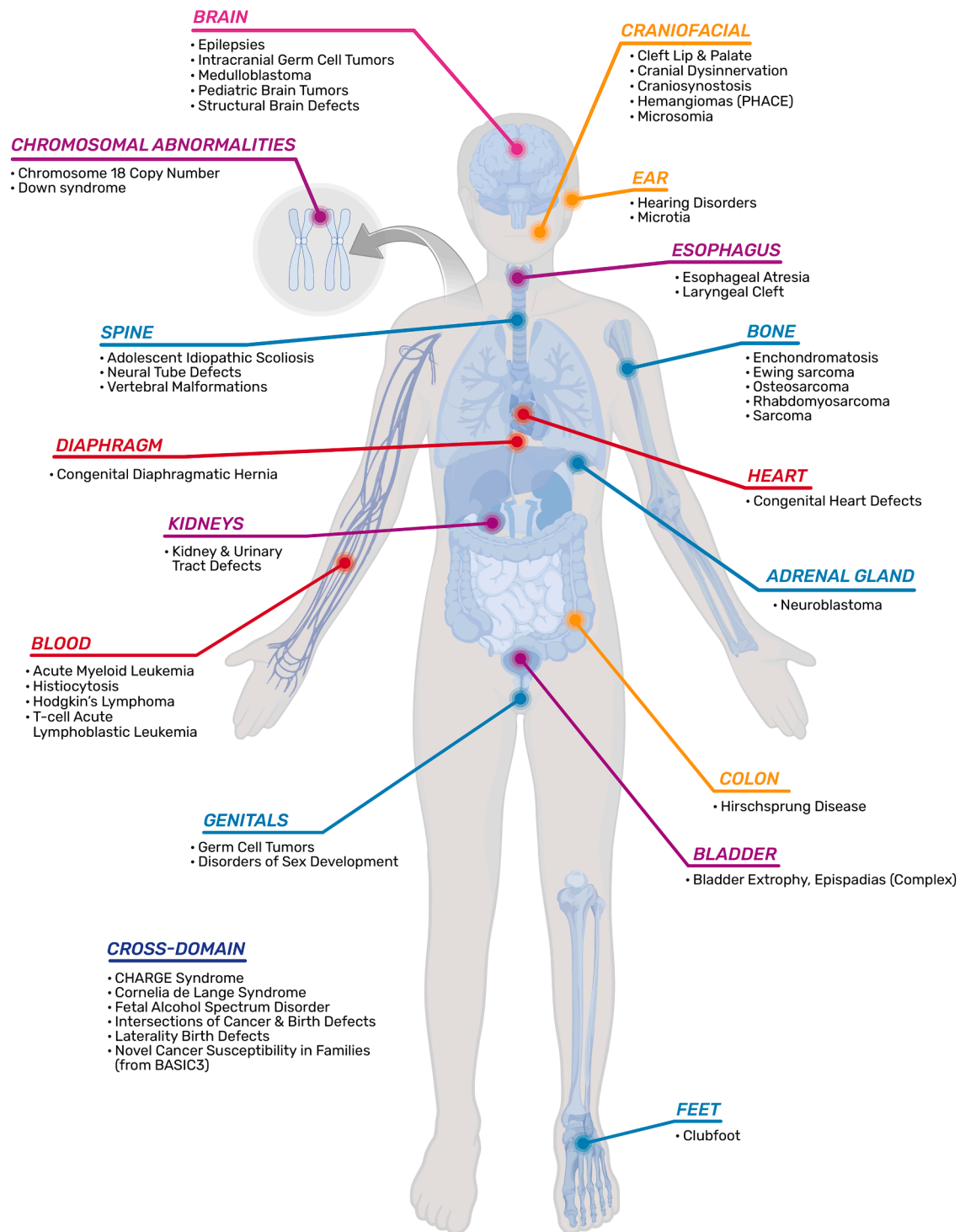


Figure 1. Disease areas of interest of the Gabriella Miller Kids First Data Resource Center

Kids First datasets are organized into studies based on conditions of interest from the submitting principal investigators. While structured in this manner, individual patients may also have conditions outside their study's original area of interest. For example, a participant enrolled in the congenital diaphragmatic hernia study may also have heart or kidney defects. The Kids First Portal's ontology search tools allow secondary investigators to query and identify participants of interest for their own research projects.

Table 1. Self-reported participant demographics of the 36 studies on the Kids First Portal

Total participants on the Kids First Portal with WGS or RNA data	30,854
Sex	
Male	16,042
Female	14,472
Not available	340
Race	
White	17,497
Asian	3,037
Black or African American	1,847
More than one race	1,580
American Indian or Alaska Native	84
Native Hawaiian or other Pacific Islander	36
Not available	6,773
Ethnicity	
Hispanic or Latino	3,834
Not Hispanic or Latino	19,238
Not available	7,782
Clinical	
Participants who are study probands	15,100
Participants with at least one HPO term	14,409
Participants with at least one MONDO term	14,114
Genomic	
Participants with WGS germline data available	30,473
Participants with WGS somatic data available	3,240
Participants with RNA-seq data available	4,090
Participants with long-read sequencing data available	88
Participants with registered-tier data available	4,794

race and ethnicity as provided by the participants and does not carry out its own genetic ancestry testing. Of the participants with genomic data for whom race and ethnicity data are available, large majorities of participants identify themselves as White (72.7%) and not Hispanic or Latino (83.3%), though some individual studies have higher proportions of minority groups, such as Kids First: Orofacial Clefts in African and Asian Ancestry and Kids First: Genetics of Microtia in Hispanic Populations.

Genomic data files on the Kids First Portal are released under one of two tiers. Register-tier files, such as somatic variants and gene expression counts, are accessible to researchers who register for the portal, with no additional permissions required. Controlled-tier files, such as aligned

reads and germline variants, represent more sensitive data and can only be accessed by a user who is approved by the Kids First Data Access Committee (DAC) through the National Center for Biotechnology Information (NCBI)'s Database of Genotypes and Phenotypes (dbGaP) Authorized Access System.¹⁹

Advances in sequencing technology, software, and genome annotations have all challenged the Kids First DRC to maintain high-quality datasets that are relevant to today's scientific community. With each new version of a bioinformatic tool, the Kids First DRC must weigh the benefits of any improvements against the costs associated with rerunning and updating existing data. The Kids First DRC has also provided annotations from databases such as ClinVar²⁰ and scores from the Combined

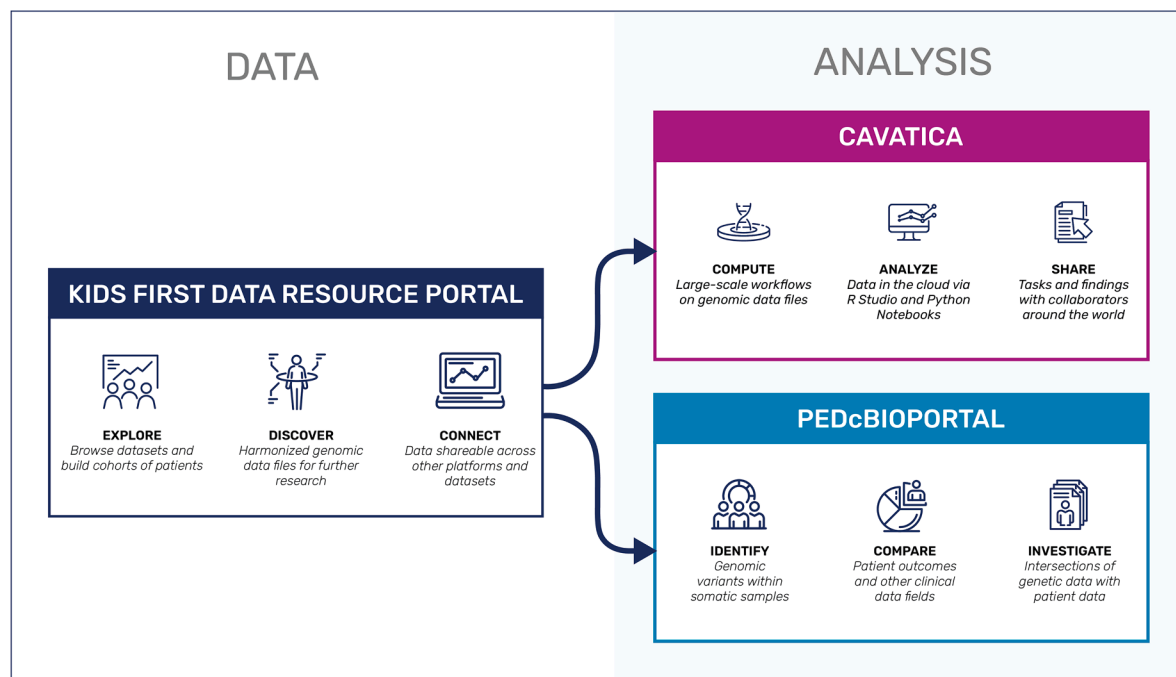


Figure 2. Platforms of the Kids First Data Resource

The Kids First Data Resource Center supports three major platforms for data discovery and analysis. Users typically start their workflow on the Kids First Portal and search for participants of interest using the data exploration filtering tool. From there, users can push data files to CAVATICA for cloud-based analysis and collaboration. Users can also analyze somatic data in the PedcBioPortal, where they can look for correlations between clinical and genomic features within Kids First's cancer studies.

Annotation Dependent Depletion (CADD)²¹ tool within its variant files, helping to support rapid variant discovery by users. These annotations go stale as new versions of annotation databases are released. One solution to addressing this issue is to support variant exploration and annotation with user-friendly tooling in our cloud-based platforms, as described in the following section.

Kids First has expanded into other modalities beyond short-read sequencing. Proteomics data from three Kids First studies are available via NCI's Clinical Proteomic Tumor Analysis Consortium (CPTAC)'s Data Portal.²² The proteomic samples are paired with short-read samples for combined, multimodal analyses. Long-read sequencing technologies from the Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) platforms outperform short-read methods in structural variant detection and tandem repeat expansion, as reviewed in the context of rare disease²³ and cancer.²⁴ The Kids First Program has selected ten cohorts for long-read sequencing across both platforms, three of which are available on the Kids First Portal today.

A limitation of Kids First clinical data is that they only represent a snapshot of an individual's condition at the time each study is awarded. Because Kids First studies originate from different institutions and are organized by different investigators, the quality and depth of phenotyping vary from study to study. Each study typically has detailed phenotype information on its condition of focus but may not be as deep in others. For example, the

Kids First: Congenital Heart Defect study has very specific cardiac phenotypes of its participants, while a participant in a cranial development study might only be described as having a "heart defect" without further detail. This uneven phenotyping can make it difficult to build a cohort of participants from multiple studies for a combined analysis. Furthermore, Kids First does not have a mechanism to update clinical data records longitudinally to track participants' health and change their records after their study has been released on the Kids First Portal. Because participants are consented under different protocols at different hospitals by different research groups, follow-ups led by the Kids First DRC are not possible, and clinical data updates are much more limited than they are in similar programs such as St. Jude Cloud and All of Us (described further in [our data resource](#)).

Our integrated platforms

Beyond the production of high-quality clinical and genomic datasets, the Kids First DRC is responsible for building and maintaining the technical infrastructure required to make these data available to approved users around the world. We support three integrated platforms: the Kids First Data Resource Portal, CAVATICA, and PedcBioPortal (Figure 2). Each of these is powered by cloud-based computing technologies equipped to address the challenges of operating a data commons, such as

managing petabytes of data and authenticating data access to users worldwide with reduced data duplication.²⁵

Cloud computing adheres to the NIH requirement for data to be consistent with the FAIR (findable, accessible, interoperable, and reusable) data principles.²⁶ As a publicly funded data resource, the Kids First DRC's platforms are built according to the FAIR principles and focused on making data "FAIR" for all users.^{26,27}

The Kids First Data Resource Portal (<https://portal.kidsfirstdrc.org>; Kids First Portal) is the main platform through which users can discover and access Kids First data. The portal's data exploration tool allows users to apply data filters to identify cohorts of participants, biospecimens, or genomic files of interest. Files within a cohort that have been selected by a user on the Kids First Portal can be pushed to a private workspace on CAVATICA (<https://cavatica.sbgenomics.com>), Kids First's cloud-based analysis platform built by Seven Bridges Genomics (a Velsera company). CAVATICA offers a repository of publicly available bioinformatic workflows suitable for data analysis, maintained by the Velsera and Kids First DRC teams. Experienced users can build their own apps and bring their own pipelines into the CAVATICA environment to run custom analyses. The CAVATICA Data Studio feature offers interactive access to private files in the project, allowing additional analysis through either RStudio or Python-based Jupyter Notebook environments.

PedcBioPortal (<https://pedcbiportal.kidsfirstdrc.org>) is an installation of the cBioPortal for Cancer Genomics,^{28,29} which enables analysis of somatic data from Kids First's cancer studies using a point-and-click interface. Users can group participants based on clinical features or somatic variants before comparing patient outcomes and identifying correlated genomic features. Because participant and biospecimen IDs are consistent across all Kids First platforms, information from one can be used to connect to the others. For example, users can define cohorts of participants based on somatic variants in PedcBioPortal, identify germline files associated with those participants in the Kids First Portal, and push those files to CAVATICA for further analysis.

All Kids First DRC platforms are built atop an Amazon Web Services (AWS) cloud-based infrastructure, with publicly available tooling in GitHub (<https://github.com/kids-first>). CAVATICA also supports Google and Azure workspaces for analysis. By avoiding a locally managed deployment of web services hosted at a member institution, the Kids First DRC has largely benefited from the commonly cited advantages of working in the cloud, including scalable storage, collaborative development across international institutions, and portable workflows to support users.^{25,30,31}

While the cloud-based design of the Kids First DRC's infrastructure allows the data resource to grow easily, some challenges still emerge. Being built atop scalable cloud infrastructure requires more careful management

of billing and costs than a system operating locally. We partnered with the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative to support the management of cloud costs. Additional challenges arise when we need to move the cloud locations of portions of the data resource. This occurred recently with a data migration from storage buckets managed by the Kids First DRC to a new set managed by NCI. It also occurred on the platform level when we changed the underlying infrastructure to draw from a new Fast Healthcare Interoperability Resources (FHIR)-based standard³² in place of our original data model. We address such challenges by leveraging APIs that facilitate the exchange of data from multiple locations, which coalesce into a final, production-level layer that users access via the interface on our platforms. These production-level data remain unchanged, while the source of the underlying data is updated without disrupting the user experience.

Our data resource

The intersection of Kids First's genomic datasets and cloud-based platforms results in the Kids First Data Resource, empowering discoveries in pediatric cancer and congenital anomalies. The Kids First DRC has worked to build tools to quickly search for variants of interest within the Kids First datasets on our platforms that are accessible to biomedical researchers of all types, including those with limited computational biology expertise.

The Variant Explorer on the Kids First Portal allows researchers to query for variants of interest using a point-and-click interface within their browser. After providing a list of genes of interest, the Variant Explorer quickly returns a table of variants in those genes identified within samples from Kids First participants. Further context about the variants is also provided, including gnomAD allele frequencies (v.3.1.2),³³ annotations from the ClinVar database,²⁰ and scores from tools such as the Variant Effect Predictor (VEP)³⁴ and CADD.²¹ Because Variant Explorer does not allow the direct identification of the individual participants with those variants, the Kids First DRC can make Variant Explorer usable by anyone who logs into the Kids First Portal, without any additional data access requirements, making Variant Explorer a useful tool for rapid searches to identify studies to apply for full access for further investigation.

The Variant Workbench on CAVATICA is a more powerful alternative to Variant Explorer. Variant Workbench is a Jupyter notebook tool within CAVATICA's interactive Data Studio environment that allows users to query Kids First variant data and annotation tools using PySpark SQL. This environment combines Kids First genetic and clinical data for more advanced queries, connecting the same variant annotations described above with the participants themselves and the health conditions they were

diagnosed with. Because it allows direct identification of individual participants' germline variants, Variant Workbench requires prior access approval for the underlying datasets via dbGaP to utilize these advanced features.

The Kids First Data Resource shares similarities with the St. Jude's Cloud Genomics Platform.³⁵ Both are large cloud-based repositories with genomic data from thousands of patients connected to analysis platforms for research and visualization. The St. Jude's Cloud features datasets from pediatric cancers, though it does not have the same amount of data from children with congenital anomalies. The St. Jude's Cloud does have longitudinal cancer survivorship data, which exceeds Kids First's current capabilities for following up with participants and presenting up-to-date clinical outcomes. Beyond the datasets, the policy and governance of the two resources are quite different. As a publicly funded data resource, access to Kids First's controlled-tier data is mediated by the Kids First DAC via dbGaP, not the individual hospitals at which Kids First participants enrolled in their respective studies.

The Kids First Data Resource is also compared to All of Us, another NIH-funded program that integrates clinical and genomic data for large-scale analyses.³⁶ All of Us is a much larger program, primarily focused on human health in unrelated adults. Kids First shares genomic data from probands as well as affected and unaffected family members but does not capture environmental and exposure data at the scale of All of Us. Kids First and All of Us have similar data modalities, including short-read WGS and long-read data that could be compatible for cross-program analysis. The All of Us limitation on bringing participant-level data out of their Researcher Workbench contrasts with Kids First, which is more broadly interoperable beyond the platforms we directly support. All of Us governance also differs from Kids First; because All of Us directly recruits and consents participants, they can draw from electronic health records and recontact participants about data they self-report. Kids First is a harmonized collection of participants enrolled in different studies at different institutions. Kids First clinical data are submitted by researchers running each trial, which allows for an additional level of curation but makes the collection and processing of subsequent clinical events on the Kids First Portal more challenging. The Kids First DRC is interested in incorporating new tooling that can unlock longitudinal updates and the introduction of new data elements for the community to research (see [our future](#)).

The Kids First Data Resource combines pediatric cancer and congenital anomaly cohorts into one harmonized collection of data for cross-study genetic discoveries. The trio-based structure of most anomaly studies allows for *de novo* and inherited variant analysis within individual families, with tumor-normal pairs allowing for somatic mutation identification within cancer cases. All of these are directly interoperable with other large publicly funded

consortia. As members of the NIH's Common Fund Data Ecosystem (CFDE),³⁷ the Kids First DRC collaborated with the Genotype-Tissue Expression (GTEx) Program³⁸ to make our transcriptomic datasets functionally equivalent, enabling researchers to combine Kids First DRC expression data with tissue-specific controls from GTEx. The Kids First DRC's cloud-based data resource allows us to efficiently partner with other large-scale cloud-based research efforts. As members of the NIH Cloud Platform Interoperability Program (NCPI),³⁹ we have enabled Kids First data to be analyzed on other bioinformatic workspaces not directly managed by Kids First DRC staff members, such as the Broad Institute's Terra platform.⁴⁰ Likewise, we have enabled users to be able to bring cloud-based data from outside Kids First into CAVATICA, such as dbGaP,¹⁹ the NCI Cancer Genomics Cloud,¹⁵ the NHLBI's BioData Catalyst,⁴¹ and the INCLUDE (investigation of co-occurring conditions across the lifespan to understand Down Syndrome) Data Hub.⁴² Connecting these platforms allows for seamless interoperability of data, empowering users to more easily combine datasets and make their own choices about where they carry out their analyses.

The Kids First DRC supports researchers looking to learn more about the platforms and how to use them in research projects, with documentation and videos outlining major Portal functions available online within the Kids First DRC Help Center (<https://kidsfirstdrc.org/help-center>). The Kids First DRC regularly participates in educational, training, and outreach events focused on using our datasets and platforms. These have included presentations and exhibits at major academic conferences, such as the American Society of Human Genetics (ASHG) Annual Meeting and the American Association for Cancer Research (AACR). The Kids First DRC hosts three public webinars each year, which typically feature presentations on research findings, new portal features, and the promotion of federal funding opportunities for research projects working on Kids First data. One webinar is focused specifically on addressing the interests of the patients and families affected by the conditions represented in our data resource, allowing for the return of information to these communities, which have been influential in the shaping of the program. All these activities serve to benefit the research community by enhancing data utilization through both the awareness of our data resource and training in how to carry out omics-level research.

Our impact

The Kids First Data Resource is actively used by genomic researchers around the world to identify the genetic causes of pediatric conditions. As of the end of 2025, the Kids First Portal has more than 5,300 registered users, while CAVATICA has more than 4,000 registered accounts. In that time period, 1,168 data-use applications

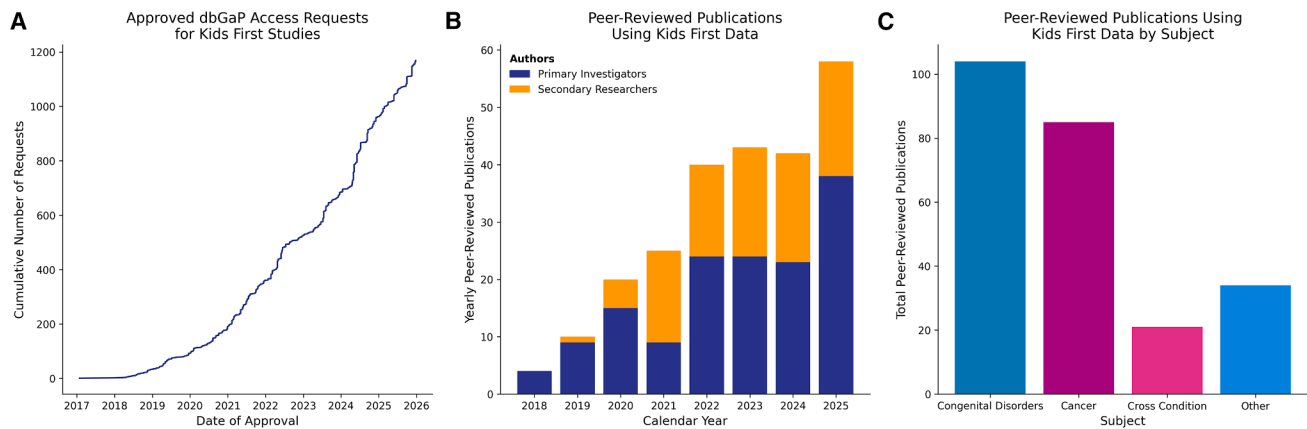


Figure 3. Impact of the Kids First Data Resource

(A) Controlled-tier data on the Kids First Portal are only accessible to users who have applied to access the study via dbGaP. As of the end of 2025, 1,168 data access requests were approved for Kids First studies.

(B) Research using Kids First datasets is published in peer-reviewed journal articles. The number of articles in each calendar year is presented for two categories of authors. Primary investigators (shown in dark purple) are the investigators who consented patients and submitted the cohorts for sequencing. These articles include at least one primary investigator as an author. Secondary researchers (shown in yellow-orange) are researchers unaffiliated with the original study and who only accessed the data via secondary use on our platforms. These articles do not include a primary investigator as an author.

(C) The same list of publications, organized by subject matter. Most of the articles focus solely on congenital disorders (104) or pediatric cancer (85) conditions. A small number investigate the intersection of both cancer and congenital disorders (21). Publications listed as other typically feature the development of a piece of bioinformatics software and use Kids First as a benchmarking dataset.

for Kids First studies have been approved via dbGaP (Figure 3A), indicating interest from the broader research community in using these data for their research.

To investigate the types of research being done using Kids First datasets, we conducted a literature search using Google Scholar. We searched for articles that include either the exact text phrase “Gabriella Miller Kids First” or one of the dbGaP accession numbers for a Kids First study, as typically found in a methods or data availability section. We manually reviewed the output to remove results that were not written articles (such as conference posters or presentations), not peer reviewed (such as pre-prints), or only mentioned Kids First in passing. To match the data-use application quantification, we limited our search to papers published in 2025 or earlier. To put Kids First into context, we also looked at publication metrics for comparator programs described above. The St. Jude Cloud directs researchers to cite their 2021 paper in *Cancer Discovery*,³⁵ which had 151 citations in PubMed during this same time period. All of Us requests that researchers who are publishing manuscripts share their publications back with the program; their website shows 1,178 publications during this time period.

We found 244 papers and categorized them by year of publication, author group, and subject. Scholarship using Kids First data has steadily increased since 2018 (Figure 3B). 60% of the publications we identified listed a principal investigator (PI) of the original Kids First study as an author. The PI is the leader of the research group responsible for consenting the patients and collecting specimens (“primary investigators”). On the contrary, 40% of the publications did not list the original PI as an

author (“secondary researchers”), demonstrating the utility of Kids First datasets not just to the original research group but for secondary analysis as well. The subject matter of the papers was broadly categorized to look for trends (Figure 3C). Most focus specifically on congenital anomalies (42.6%) or pediatric cancer (34.8%), with a smaller amount looking across both congenital anomalies and cancer (8.6%). Of note, a large percentage of manuscripts we categorized as “other” (13.9%); these are largely papers focused more on the development of bioinformatics software tooling and less on research into specific biological conditions, whereas the Kids First dataset is used as a large, harmonized collection of genomes for testing and benchmarking. Below, we summarize some of the findings made by researchers using Kids First data. Examples of use of Kids First pediatric cancer data have been discussed previously.⁴³

Kids First datasets have been used to support genome-wide association studies (GWASs), including gene discovery experiments investigating orofacial clefts,⁴⁴ development of the face and ear,⁴⁵ and craniosynostosis.⁴⁶ Because Kids First datasets are harmonized using the same bioinformatic workflows, datasets can be easily joined together for a combined analysis to enhance statistical power and increase the likelihood of discoveries. This is demonstrated through the combination of multiple Kids First orofacial cleft studies,⁴⁷ the Kids First Microtia and Craniofacial Microsomia study,⁴⁸ and by using the Kids First: Ewing Sarcoma study to validate findings in cleft lip/palate data.⁴⁹

The enhanced interoperability of Kids First datasets and platforms facilitates the combination of Kids First data

with those of other major consortia, allowing for more advanced cross-disease analyses. Foss-Skiftesvik et al. identified that variants in *CDKN2B-AS1* increased the likelihood of developing pediatric gliomas by combining Kids First data with other pediatric cancer datasets, including the Glioma International Case-Control Study and validation using data from GTEx.⁵⁰ Nabbi et al. combined Kids First pediatric brain tumor data with datasets from TCGA and TARGET programs to identify immune responses to nervous system tumors at the transcriptional level.⁵¹ Spendlove et al. used publicly available summary statistics from GWASs on the UK Biobank cohort to calculate polygenic risk scores associated with congenital heart defects, which were then tested against the Kids First: Congenital Heart Defects study.⁵² Investigators also combined Kids First datasets with those of their own research groups, as was used to identify the role of variants in DNA damage repair genes in predisposition to Ewing sarcoma⁵³ and non-coding *de novo* variants in congenital heart disease.⁵⁴

Kids First datasets also serve as a readily available resource for answering research questions outside the scope of variant discovery. Investigators can bring their own expertise to the harmonized genomic files and rapidly execute their own custom workflows on Kids First data. For example, Kaneva et al. measured mitochondrial ancestry in the Kids First: Ewing Sarcoma dataset to study the role of the mitochondrial genome in the correlation of the incidence of Ewing sarcoma and race.⁵⁵ Kids First data have been used in the development of bioinformatic software.⁵⁶ The trio-based structure of many studies allowed for the specialized calling of compound heterozygous variants across multiple Kids First cohorts.⁵⁷ Kids First data have also been combined with work in model systems, validating human variants associated with esophageal atresia in *Xenopus*,⁵⁸ human variants associated with craniofacial development in zebrafish,⁵⁹ and human enhancers in variants associated with congenital heart defects in human induced pluripotent stem cell-derived cardiomyocytes.⁶⁰ Researchers have also worked in the opposite direction, investigating the larger role of a gene that affects heart development in mice by searching for variants in Kids First's human data.⁶¹

While Kids First data are used to make discoveries for individual conditions, one of the goals of the program is to find shared underlying genetic pathways between pediatric cancer and congenital anomalies. A 2025 study used Kids First data to identify seven genes with mutations enriched in both congenital heart defect and neuroblastoma cohorts, supporting a shared developmental mechanism for these conditions.⁶² A different study from the same year compared copy-number variants in Kids First data from patients with cancer and congenital anomalies vs. those with congenital anomalies alone; their results found a high degree of heterogeneity at the gene level but more commonalities at the pathway level, with genes associated with the synapse organization pathway being enriched in

patients with both cancer and a congenital anomaly.⁶³ We are excited by these findings, but overall, the amount of scholarship in this “cross-condition” research has lagged behind that of the individual conditions (Figure 3C). We are hopeful that Kids First datasets can be used to further investigate these pathways and find new ones.

The most impactful findings using Kids First data are those that can benefit the health of pediatric patients. Specific genomic variants in the Kids First congenital diaphragmatic hernia,⁶⁴ neuroblastoma,⁶⁵ and T cell acute lymphoblastic leukemia cohorts were identified to be associated with worse outcomes for patients diagnosed with these conditions.⁶⁶ Screening for these variants upon diagnosis can identify more severe cases earlier, directing better treatments from clinical care teams. Kids First datasets were also used to identify risk factors in the genome that predispose individuals to neuroblastoma⁶⁷ and lymphoma,⁶⁸ thus empowering better screening and early detection methods. Finally, the large collection of Kids First data has been used to identify molecular subtypes that can categorize or subtype diagnoses, as was done by Li et al. for children diagnosed with acute lymphoblastic leukemia and Down syndrome,⁶⁹ enabling targeted therapies and treatments.

Our future

The Gabriella Miller Kids First DRC was founded on the goal to identify the genetic causes of and links between pediatric cancer and congenital anomalies. We believe we have achieved success toward this goal, as seen in the scientific literature, and the synthesis of datasets and platforms into the Kids First Data Resource will continue to allow future researchers to make further discoveries in the future. At the same time, true success would represent a world where patients with these conditions have improved treatment options, better outcomes, and longer, happier lives with their families. As we enter the second decade since the passage of the Gabriella Miller Kids First Research Act, we look to continue building the Kids First Data Resource while acknowledging that the legacy of the Kids First DRC will lie not with the academic scholarship but with how these resources and discoveries impact patients in the clinic.

In the next calendar year, the Kids First DRC is planning several major data releases to significantly expand and enhance the Kids First Data Resource. These include a re-release of every Kids First germline WGS genomic file with a new output generated using Illumina's DRAGEN bioinformatics tooling, which has improved runtimes and accuracy compared to the GATK-based tooling our pipelines are currently based on.⁷⁰ We anticipate uploading thousands of additional short-read genomes, expanding existing studies, and releasing entirely new datasets. In addition to releasing the genomic files for these, we will also make the variants available in our Variant Explorer

and Variant Workbench tools along with up-to-date annotations such as gnomAD frequencies. Beyond short-read data, we will enhance our long-read data with hundreds of additional long-read samples. Methylation data will also be made available, a first for the Kids First Portal. These user-facing improvements are powered by upgrades to our underlying clinical data infrastructure and a new data workflow developed in partnership with NCI's CCDI.

Beyond the next 12 months, the Kids First DRC is researching ways to better integrate our database with the active medical records of the participants represented on our platforms. At present, clinical information presented on the Kids First Portal only shares a snapshot in time of participants' health. As highlighted above, Kids First's clinical data are not as robust as similar large-scale efforts such as the St. Jude's Cloud or the All of Us Research Program, without a longitudinal element to track participants through their treatment journey. Better incorporation of ongoing information about how participants receive treatment and are diagnosed with additional conditions will expand the types of research that can be done using our data resource and ensure that any analyses have the most up-to-date clinical information. Just as important as these new data modalities is building technical infrastructure that will allow the return of knowledge to the clinic—connecting information about the genetics, treatment, and outcomes of Kids First participants with clinicians worldwide who can use the platform to search for similar cases to those of the patients they are caring for.

To achieve these goals, we will explore better ways to automate and extract clinical data from electronic health records. One of the biggest bottlenecks in the clinical research space is the process of reading information and transforming it into other formats for research uses. This labor-intensive work is slow and expensive to manually review records and extract specific pieces of data for downstream projects. Automating this work will increase the amount of data available and unlock longitudinal updates for Kids First participants with near-real-time extraction of information. This acceleration will not only improve research quality but will also improve the return of research findings to guide care. Clinicians will be able to review more data in one location to find participants similar to their patients to inform clinical trial outcomes and treatment decisions. One such initiative is the Pediatric Care Expansion (PCX), funded by the Advanced Research Projects Agency for Health (ARPA-H). We are exploring the use of their tooling to demonstrate how better clinical data integration can improve Kids First projects in the future.

Precision medicine offers a great promise of improved patient health as we enter the middle of the century with big data in space and time. The Kids First DRC's expertise in data generation and platform development uniquely positions it to address the technical challenges associated with augmenting pediatric care at scale, with much more work needed. We conclude as Gabriella urged us, choosing to “stop talking” and get back to the “doing” that lies ahead.

Acknowledgments

This work is supported by funding from the NIH, grant numbers U2CHL138346, U2CHD109731, and OT2OD030162. Writing and editing support was provided by Darby Major, writer/editor, Division of Neurosurgery at the Children's Hospital of Philadelphia. We would like to acknowledge the contributions of all our NIH colleagues who have worked with the Kids First DRC, including James Coulombe, Jonathan Kaltman, Kathryn Stein, Lorette Javois, and Maria Concepcion Nierras. We acknowledge our partners at Kids First sequencing centers, including the Broad Institute and HudsonAlpha Institute for Biotechnology, as well as X01 investigators, our domain-expert partners who agree to share data with the broader research community. Finally, we thank and acknowledge the contributions of the patients and their families who enrolled in clinical research and contributed samples for sequencing, which ultimately became Kids First cohorts—each of them share Gabriella Miller's spirit, fight, and dream of a future without sick children.

Author contributions

The manuscript was written by D.M.H. with editing and suggestions by E.B., V.C., D.D., M.V.F., A.C.H., H.L., R.C., J.D., and A.P.H. The study was conceived and designed by M.K., B.D.-D., M.A.H., J.L., P.B.S., S.V., R.L.G., R.C., J.D., D.T., V.F., A.R., and A.P.H. Clinical data were reviewed and curated by C.D., C.F., N.S., A.W., B.F., N.A.V., and A.P.H. The bioinformatic pipelines were produced and maintained by M.B., Y.G., Q.L., D.M., A.S., J.W., B.Z., C.Z., and Y.Z. The Kids First Portal and its underlying infrastructure were built and maintained by C.B., A.L., G.A., D.B., J.C., E.G., L.-F.L., L.L., A.P., C.P., K.S.-O., J.-P.T., and V.F. CAVATICA and its underlying infrastructure were built and maintained by M.M., S.S., J.R., M.T., M.Z., B.D.-D., and J.D. The Gen3 software used for data access authentication was built and maintained by J.K., G.K., R.R., P.R., S.V., and R.L.G. The underlying Kids First data model was built and maintained by N.S., Y.C., A.E., J.F.V., S.F., S.L., E.T., K.Y.H., R.C., and A.P.H. Administration and project management were provided by D.M.H., L.D., B.M., W.P., E.R., W.R., E.W., J.-P.T., B.F., E.B., V.C., D.D., M.V.F., J.M.G.-A., A.C.H., H.L., and L.W.

Declaration of interests

M.A.H. is a founder of Alamy Health.

Web resources

CAVATICA, <https://cavatica.sbgenomics.com>
 Gabriella Miller Kids First Data Resource Portal, <https://portal.kidsfirstdrc.org>
 PedcBioPortal, <https://pedcbiportal.kidsfirstdrc.org>

References

1. Curtin, S.C., Tejada-Vera, B., Bastian, B.A.; and National Center for Health Statistics (2024). Deaths: Leading Causes for 2022. *Natl. Vital Stat. Rep.* 73, 2024.
2. Howlader, N., Noone, A.-M., Krapcho, M., Garshell, J., Neyman, N., Altekruse, S.F., Kosary, C.L., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., et al. (2014). SEER Cancer Statistics Review,

- 1975–2012 (National Cancer Institute). https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/3420543.
3. Sweet-Cordero, E.A., and Biegel, J.A. (2019). The genomic landscape of pediatric cancers: Implications for diagnosis and treatment. *Science* 363, 1170–1175. <https://doi.org/10.1126/science.aaw3535>.
4. Centers for Disease Control and Prevention CDC (2008). Update on overall prevalence of major birth defects—Atlanta, Georgia, 1978–2005. *MMWR Morb. Mortal. Wkly. Rep.* 57, 1–5.
5. Stallings, E.B., Isenburg, J.L., Rutkowski, R.E., Kirby, R.S., Nembhard, W.N., Sandidge, T., Villavicencio, S., Nguyen, H.H., McMahon, D.M., Nestoridi, E., et al. (2024). National population-based estimates for major birth defects, 2016–2020. *Birth Defects Res.* 116, e2301. <https://doi.org/10.1002/bdr2.2301>.
6. Glinianaia, S.V., Embleton, N.D., and Rankin, J. (2012). A systematic review of studies of quality of life in children and adults with selected congenital anomalies. *Birth Defects Res. A Clin. Mol. Teratol.* 94, 511–520. <https://doi.org/10.1002/bdra.23030>.
7. Garcia Rodrigues, M., Monteiro Soares, M., Rodrigues, J.D., Azevedo, L.F., Pereira Rodrigues, P., Areias, J.C., and Areias, M.E. (2022). Quality of life of parents with children with congenital abnormalities: a systematic review with meta-analysis of assessment methods and levels of quality of life. *Qual. Life Res.* 31, 991–1011. <https://doi.org/10.1007/s11136-021-02986-z>.
8. Robinson, K., Parrish, R., Adeyemo, W.L., Beaty, T.H., Butali, A., Buxó, C.J., Gowans, L.J.J., Hecht, J.T., Moreno Uribe, L., Murray, J.C., et al. (2024). Genome-wide study of gene-by-sex interactions identifies risks for cleft palate. *Hum. Genet.* 143, 1341–1352. <https://doi.org/10.1007/s00439-024-02704-y>.
9. Wynn, J., Yu, L., and Chung, W.K. (2014). Genetic causes of congenital diaphragmatic hernia. *Semin. Fetal Neonatal Med.* 19, 324–330. <https://doi.org/10.1016/j.siny.2014.09.003>.
10. Schraw, J.M., Desrosiers, T.A., Nembhard, W.N., Langlois, P.H., Meyer, R.E., Canfield, M.A., Rasmussen, S.A., Chambers, T.M., Spector, L.G., Plon, S.E., and Lupo, P.J. (2020). Cancer Diagnostic Profile in Children with Structural Birth Defects: An Assessment in 15,000 Childhood Cancer Cases. *Cancer* 126, 3483–3492. <https://doi.org/10.1002/cncr.32982>.
11. Lupo, P.J., Schraw, J.M., Desrosiers, T.A., Nembhard, W.N., Langlois, P.H., Canfield, M.A., Copeland, G., Meyer, R.E., Brown, A.L., Chambers, T.M., et al. (2019). Association Between Birth Defects and Cancer Risk Among Children and Adolescents in a Population-Based Assessment of 10 Million Live Births. *JAMA Oncol.* 5, 1150–1158. <https://doi.org/10.1001/jamaoncol.2019.1215>.
12. Johnson, K.J., Lee, J.M., Ahsan, K., Padda, H., Feng, Q., Parthap, S., Fowler, S.A., and Druley, T.E. (2017). Pediatric cancer risk in association with birth defects: A systematic review. *PLoS One* 12, e0181246. <https://doi.org/10.1371/journal.pone.0181246>.
13. Baugh, J.N., Veldhuijzen van Zanten, S., Fiocco, M., Colditz, N., Hoffmann, M., Janssens, G.O., Valentini, C., Hargrave, D., Wiese, M., von Bueren, A.O., et al. (2024). Treatment-related survival patterns in diffuse intrinsic pontine glioma using a historical cohort: A report from the European Society for Pediatric Oncology DIPG/DMG Registry. *Neurooncol. Adv.* 6, vdae155. <https://doi.org/10.1093/noajnl/vdae155>.
14. Janssens, G.O., Gandola, L., Bolle, S., Mandeville, H., Ramos-Albiac, M., van Beek, K., Benghiat, H., Hoebe, B., Morales La Madrid, A., Kortmann, R.D., et al. (2017). Survival benefit for patients with diffuse intrinsic pontine glioma (DIPG) undergoing re-irradiation at first progression: A matched-cohort analysis on behalf of the SIOP-E-HGG/DIPG working group. *Eur. J. Cancer* 73, 38–47. <https://doi.org/10.1016/j.ejca.2016.12.007>.
15. Jensen, M.A., Ferretti, V., Grossman, R.L., and Staudt, L.M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. *Blood* 130, 453–459. <https://doi.org/10.1182/blood-2017-03-735654>.
16. National Cancer Institute (2025). Therapeutically Applicable Research to Generate Effective Treatments (TARGET) (National Cancer Institute). <https://www.cancer.gov/ccg/research/genome-sequencing/target>.
17. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* 83, 610–615. <https://doi.org/10.1016/j.ajhg.2008.09.017>.
18. Vasilevsky, N.A., Toro, S., Matentzoglou, N., Flack, J.E., Mullen, K.R., Hegde, H., Gehrke, S., Whetzel, P.L., Shwetar, Y., Harris, N.L., et al. (2026). Mondo: Integrating Disease Terminology Across Communities. *Genetics* 232, iyaf215. <https://doi.org/10.1093/genetics/iyaf215>.
19. Tryka, K.A., Hao, L., Sturcke, A., Jin, Y., Wang, Z.Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M., and Feolo, M. (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* 42, D975–D979. <https://doi.org/10.1093/nar/gkt1211>.
20. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985. <https://doi.org/10.1093/nar/gkt1113>.
21. Schubach, M., Maass, T., Nazaretyan, L., Röner, S., and Kircher, M. (2024). CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res.* 52, D1143–D1154. <https://doi.org/10.1093/nar/gkad989>.
22. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., and Ketchum, K.A. (2015). The CPTAC Data Portal: a resource for cancer proteomics research. *J. Proteome Res.* 14, 2707–2713. <https://doi.org/10.1021/pr501254j>.
23. Del Gobbo, G.F., and Boycott, K.M. (2025). The additional diagnostic yield of long-read sequencing in undiagnosed rare diseases. *Genome Res.* 35, 559–571. <https://doi.org/10.1101/gr.279970.124>.
24. Li, Q., Keskus, A.G., Wagner, J., Izydorczyk, M.B., Timp, W., Sedlazeck, F.J., Klein, A.P., Zook, J.M., Kolmogorov, M., and Schatz, M.C. (2025). Unraveling the hidden complexity of cancer through long-read sequencing. *Genome Res.* 35, 599–620. <https://doi.org/10.1101/gr.280041.124>.
25. Grossman, R.L. (2019). Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data. *Trends Genet.* 35, 223–234. <https://doi.org/10.1016/j.tig.2018.12.006>.
26. Wilkinson, M.D., Dumontier, M., and Aalbersberg, I.J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018. <https://doi.org/10.1038/sdata.2016.18>.

27. (2018). NIH Strategic Plan for Data Science (National Institutes of Health). https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf.
28. Cerami, E., Gao, J., and Dogrusoz, U. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data [published correction appears in *Cancer Discov.* 2(10):960]. *Cancer Discov.* 2, 401–404. <https://doi.org/10.1158/2159-8290.CD-12-0095>.
29. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6, pii. <https://doi.org/10.1126/scisignal.2004088>.
30. Lau, J.W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., Groves-Kirkby, N., Mihajlovic, A., DiGiovanna, J., Srdic, M., et al. (2017). The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. *Cancer Res.* 77, e3–e6. <https://doi.org/10.1158/0008-5472.CAN-17-0387>.
31. Schatz, M.C., Philippakis, A.A., Afgan, E., Banks, E., Carey, V.J., Carroll, R.J., Culotti, A., Ellrott, K., Goecks, J., Grossman, R.L., et al. (2022). Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genom.* 2, 100085. <https://doi.org/10.1016/j.xgen.2021.100085>.
32. Bender, D., and Sartipi, K. (2013). HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal*, pp. 326–331. <https://doi.org/10.1109/CBMS.2013.6627810>.
33. Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alföldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 625, 92–100. <https://doi.org/10.1038/s41586-023-06045-0>.
34. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122. <https://doi.org/10.1186/s13059-016-0974-4>.
35. McLeod, C., Gout, A.M., Zhou, X., Thrasher, A., Rahbarinia, D., Brady, S.W., Macias, M., Birch, K., Finkelstein, D., Sunny, J., et al. (2021). St. Jude Cloud: A Pediatric Cancer Genomic Data-Sharing Ecosystem. *Cancer Discov.* 11, 1082–1099. <https://doi.org/10.1158/2159-8290.CD-20-1230>.
36. All of Us Research Program Investigators, Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., and Dishman, E. (2019). The “All of Us” Research Program. *N. Engl. J. Med.* 381, 668–676. <https://doi.org/10.1056/NEJMSr1809937>.
37. Charbonneau, A.L., Brady, A., Czajkowski, K., Aluvathingal, J., Canchi, S., Carter, R., Chard, K., Clarke, D.J.B., Crabtree, J., Creasy, H.H., et al. (2022). Making Common Fund data more findable: catalyzing a data ecosystem. *GigaScience* 11, giac105. <https://doi.org/10.1093/gigascience/giac105>.
38. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science.* 369, 1318–1330. <https://doi.org/10.1126/science.aaz1776>.
39. National Institutes of Health Office of Data Science Strategy. NIH Cloud Platform Interoperability Program. National Institutes of Health. Updated December 30, 2025. Accessed July 15, 2026. <https://datascience.nih.gov/nih-cloud-platform-interoperability>.
40. Terra. Terra: Science at Scale. Broad Institute of MIT and Harvard. Accessed July 15, 2026. <https://terra.bio/>.
41. Ahalt, S., Avillach, P., Boyles, R., Bradford, K., Cox, S., Davis-Dusenbery, B., Grossman, R.L., Krishnamurthy, A., Manning, A., Paten, B., et al. (2023). Building a collaborative cloud platform to accelerate heart, lung, blood, and sleep research. *J. Am. Med. Inform. Assoc.* 30, 1293–1300. <https://doi.org/10.1093/jamia/ocad048>.
42. Bardhan, S., Li, H., Tarver, E., Schramm, C., Brown, M., Garcia, L., Schwartz, B., Mazzucco, A., Natarajan, N., Walsh, E., et al. (2024). The National Institutes of Health INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndrome (INCLUDE) Project: Accelerating research discoveries for people with Down syndrome across the lifespan. *Am. J. Med. Genet. C Semin. Med. Genet.* 196, e32081. <https://doi.org/10.1002/ajmg.c.32081>.
43. Hudson, A., Fournier, M., Coulombe, J., and Dae, D. (2023). Using existing pediatric cancer data from the Gabriella Miller Kids First Data Resource Program. *JNCI Cancer Spectr.* 7, pkad079. <https://doi.org/10.1093/jncics/pkad079>.
44. Mukhopadhyay, N., Bishop, M., Mortillo, M., Chopra, P., Hetmanski, J.B., Taub, M.A., Moreno, L.M., Valencia-Ramirez, L.C., Restrepo, C., Wehby, G.L., et al. (2020). Whole genome sequencing of orofacial cleft trios from the Gabriella Miller Kids First Pediatric Research Consortium identifies a new locus on chromosome 21. *Hum. Genet.* 139, 215–226. <https://doi.org/10.1007/s00439-019-02099-1>.
45. Quia, D., Kim, S.W., Zhang, Q., Morton, S.U., Pereira, A.C., DePalma, S.R., Willcox, J.A.L., McDonough, B., DeLaughter, D.M., Gorham, J.M., et al. (2022). An ancient founder mutation located between ROBO1 and ROBO2 is responsible for increased microtia risk in Amerindigenous populations. *Proc. Natl. Acad. Sci. USA* 119, e2203928119. <https://doi.org/10.1073/pnas.2203928119>.
46. Goovaerts, S., Hoskens, H., Eller, R.J., Herrick, N., Musolf, A.M., Justice, C.M., Yuan, M., Naqvi, S., Lee, M.K., Vandermeulen, D., et al. (2023). Joint multi-ancestry and admixed GWAS reveals the complex genetics behind human cranial vault shape. *Nat. Commun.* 14, 7436. <https://doi.org/10.1038/s41467-023-43237-8>.
47. Bishop, M.R., Diaz Perez, K.K., Sun, M., Ho, S., Chopra, P., Mukhopadhyay, N., Hetmanski, J.B., Taub, M.A., Moreno-Urbe, L.M., Valencia-Ramirez, L.C., et al. (2020). Genome-wide Enrichment of De Novo Coding Mutations in Orofacial Cleft Trios. *Am. J. Hum. Genet.* 107, 124–136. <https://doi.org/10.1016/j.ajhg.2020.05.018>.
48. Quia, D., Timberlake, A.T., Curran, J.J., Cunningham, M.L., McDonough, B., Artunduaga, M.A., DePalma, S.R., Duenas-Roque, M.M., Gorham, J.M., Gustafson, J.A., et al. (2023). Damaging variants in FOXI3 cause microtia and craniofacial microsomia. *Genet. Med.* 25, 143–150. <https://doi.org/10.1016/j.gim.2022.09.005>.
49. Zieger, H.K., Weinhold, L., Schmidt, A., Holtgrewe, M., Juranek, S.A., Siewert, A., Scheer, A.B., Thieme, F., Mangold, E., Ishorst, N., et al. (2023). Prioritization of non-coding elements involved in non-syndromic cleft lip with/without cleft palate through genome-wide analysis of de novo mutations. *HGG Adv.* 4, 100166. <https://doi.org/10.1016/j.xhgg.2022.100166>.
50. Foss-Skiftesvik, J., Li, S., Rosenbaum, A., Hagen, C.M., Stoltze, U.K., Ljungqvist, S., Hjalmar, U., Schmieglow, K., Morimoto, L., de Smith, A.J., et al. (2023). Multi-ancestry

- genome-wide association study of 4069 children with glioma identifies 9p21.3 risk locus. *Neuro Oncol.* 25, 1709–1720. <https://doi.org/10.1093/neuonc/noad042>.
51. Nabbi, A., Beck, P., Delaidelli, A., Oldridge, D.A., Sudhama, S., Zhu, K., Yang, S.Y.C., Mulder, D.T., Bruce, J.P., Paulson, J.N., et al. (2023). Transcriptional immunogenomic analysis reveals distinct immunological clusters in paediatric nervous system tumours. *Genome Med.* 15, 67. <https://doi.org/10.1186/s13073-023-01219-x>.
52. Spendlove, S.J., Bondhus, L., Lluri, G., Sul, J.H., and Arboleda, V.A. (2022). Polygenic risk scores of endo-phenotypes identify the effect of genetic background in congenital heart disease. *HGG Adv.* 3, 100112. <https://doi.org/10.1016/j.xhgg.2022.100112>.
53. Gillani, R., Camp, S.Y., Han, S., Jones, J.K., Chu, H., O'Brien, S., Young, E.L., Hayes, L., Mitchell, G., Fowler, T., et al. (2022). Germline predisposition to pediatric Ewing sarcoma is characterized by inherited pathogenic variants in DNA damage repair genes. *Am. J. Hum. Genet.* 109, 1026–1037. <https://doi.org/10.1016/j.ajhg.2022.04.007>.
54. Richter, F., Morton, S.U., Kim, S.W., Kitaygorodsky, A., Wasson, L.K., Chen, K.M., Zhou, J., Qi, H., Patel, N., DePalma, S.R., et al. (2020). Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nat. Genet.* 52, 769–777. <https://doi.org/10.1038/s41588-020-0652-z>.
55. Kaneva, K., Schurr, T.G., Tatarinova, T.V., Buckley, J., Merkurjev, D., Triska, P., Liu, X., Done, J., Maglinte, D.T., Deapen, D., et al. (2022). Mitochondrial DNA haplogroup, genetic ancestry, and susceptibility to Ewing sarcoma. *Mitochondrion* 67, 6–14. <https://doi.org/10.1016/j.mito.2022.09.002>.
56. Miller, D.B., and Piccolo, S.R. (2021). trioPhaser: using Mendelian inheritance logic to improve genomic phasing of trios. *BMC Bioinf.* 22, 559. <https://doi.org/10.1186/s12859-021-04470-4>.
57. Miller, D.B., and Piccolo, S.R. (2021). A Survey of Compound Heterozygous Variants in Pediatric Cancers and Structural Birth Defects. *Front. Genet.* 12, 640242. <https://doi.org/10.3389/fgene.2021.640242>.
58. Zhong, G., Ahimaz, P., Edwards, N.A., Hagen, J.J., Faure, C., Lu, Q., Kingma, P., Middlesworth, W., Khlevner, J., El Fiky, M., et al. (2022). Identification and validation of candidate risk genes in endocytic vesicular trafficking associated with esophageal atresia and tracheoesophageal fistulas [published correction appears in *HGG*. *HGG Adv* 3, 100107. <https://doi.org/10.1016/j.xhgg.2022.100107>.
59. Caetano da Silva, C., Macias Trevino, C., Mitchell, J., Murali, H., Tsimbal, C., Dalessandro, E., Carroll, S.H., Kochhar, S., Curtis, S.W., Cheng, C.H.E., et al. (2024). Functional analysis of ESRP1/2 gene variants and CTNND1 isoforms in orofacial cleft pathogenesis. *Commun. Biol.* 7, 1040. <https://doi.org/10.1038/s42003-024-06715-3>.
60. Xiao, F., Zhang, X., Morton, S.U., Kim, S.W., Fan, Y., Gorham, J.M., Zhang, H., Berkson, P.J., Mazumdar, N., Cao, Y., et al. (2024). Functional dissection of human cardiac enhancers and noncoding de novo variants in congenital heart disease. *Nat. Genet.* 56, 420–430. <https://doi.org/10.1038/s41588-024-01669-y>.
61. Jamet, S., Ha, S., Ho, T.H., Houghtaling, S., Timms, A., Yu, K., Paquette, A., Maga, A.M., Greene, N.D.E., and Beier, D.R. (2022). The arginine methyltransferase Carm1 is necessary for heart development. *G3 (Bethesda)* 12, jkac155. <https://doi.org/10.1093/g3journal/jkac155>.
62. Tark, J.Y., Renwick, A., Yu, Y., Sabo, A., Olshan, A.F., Plon, S.E., Huff, C.D., and Lupo, P.J. (2026). Evaluating the Genetic Overlap Between Congenital Heart Disease and Neuroblastoma Risk. *Pediatr. Blood Cancer* 73, e70055. <https://doi.org/10.1002/1545-5017.70055>.
63. Liu, Y., Glessner, J., Qu, H.Q., Chang, X., Qiu, H., Wang, T., Mentch, F.D., and Hakonarson, H. (2025). Copy number variations contribute to malignant tumor development in children with serious birth defects. *Mol. Oncol.* 19, 899–912. <https://doi.org/10.1002/1878-0261.13718>.
64. Qiao, L., Wynn, J., Yu, L., Hernan, R., Zhou, X., Duron, V., Aspelund, G., Farkouh-Karoleski, C., Zygmunt, A., Krishnan, U.S., et al. (2020). Likely damaging de novo variants in congenital diaphragmatic hernia patients are associated with worse clinical outcomes. *Genet. Med.* 22, 2020–2028. <https://doi.org/10.1038/s41436-020-0908-0>.
65. Kim, J., Vaksman, Z., Egolf, L.E., Kaufman, R., Evans, J.P., Konkrite, K.L., Danesh, A., Lopez, G., Randall, M.P., Dent, M.H., et al. (2024). Germline pathogenic variants in neuroblastoma patients are enriched in BARD1 and predict worse survival. *J. Natl. Cancer Inst.* 116, 149–159. <https://doi.org/10.1093/jnci/djad183>.
66. Pölönen, P., Di Giacomo, D., Seffernick, A.E., Elsayed, A., Kimura, S., Benini, F., Montefiori, L.E., Wood, B.L., Xu, J., Chen, C., et al. (2024). The genomic basis of childhood T-lineage acute lymphoblastic leukaemia. *Nature* 632, 1082–1091. <https://doi.org/10.1038/s41586-024-07807-0>.
67. Egolf, L.E., Vaksman, Z., Lopez, G., Rokita, J.L., Modi, A., Basta, P.V., Hakonarson, H., Olshan, A.F., and Diskin, S.J. (2019). Germline 16p11.2 Microdeletion Predisposes to Neuroblastoma. *Am. J. Hum. Genet.* 105, 658–668. <https://doi.org/10.1016/j.ajhg.2019.07.020>.
68. Flerlage, J.E., Myers, J.R., Maciaszek, J.L., Oak, N., Rashkin, S.R., Hui, Y., Wang, Y.D., Chen, W., Wu, G., Chang, T.C., et al. (2023). Discovery of novel predisposing coding and non-coding variants in familial Hodgkin lymphoma. *Blood* 141, 1293–1307. <https://doi.org/10.1182/blood.2022016056>.
69. Li, Z., Chang, T.C., Junco, J.J., Devidas, M., Li, Y., Yang, W., Huang, X., Hedges, D.J., Cheng, Z., Shago, M., et al. (2023). Genomic landscape of Down syndrome-associated acute lymphoblastic leukemia. *Blood* 142, 172–184. <https://doi.org/10.1182/blood.2023019765>.
70. Behera, S., Catreux, S., Rossi, M., Truong, S., Huang, Z., Ruehle, M., Visvanath, A., Parnaby, G., Roddey, C., Onuchic, V., et al. (2025). Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat. Biotechnol.* 43, 1177–1191. <https://doi.org/10.1038/s41587-024-02382-1>.