












Making the case for an International Childhood Cancer Data Partnership

Gonçalo Forjaz , DVM, MSc¹, Betsy Kohler , MPH², Michel P. Coleman , MD³, Eva Steliarova-Foucher , PhD⁴, Serban Negoita , MD, DrPH^{1,5,*}, Jaime M. Guidry Auvil , PhD^{1,6}, Fernanda Silva Michels , MSc, PhD², Johanna Goderre , MPH^{1,5}, Charles Wiggins , PhD⁷, Eric B. Durbin , DrPH⁸, Gijs Geleijnse, PhD⁹, Marie-Charlotte Henrion, MSc¹⁰, Candice Altmayer , MSc¹⁰, Thomas Dubois, PhD¹⁰, Lynne Penberthy, MD, MPH^{1,5}

¹Public Health Practice, Westat, Inc., Rockville, MD 20850, United States

²North American Association of Central Cancer Registries, Springfield, IL 62704, United States

³Cancer Survival Group, London School of Hygiene & Tropical Medicine, London, United Kingdom

⁴Cancer Surveillance Branch, International Agency for Research on Cancer, Lyon, France

⁵Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, MD 20850, United States

⁶Center for Biomedical Informatics & Information Technology, National Cancer Institute, Rockville, MD 20850, United States

⁷New Mexico Tumor Registry, University of New Mexico Comprehensive Cancer Center, Albuquerque, NM 87131, United States

⁸Kentucky Cancer Registry, Markey Cancer Center, Lexington, KY 40504, United States

⁹Netherlands Comprehensive Cancer Organisation, Utrecht, The Netherlands

¹⁰French National Cancer Institute, Paris, France

*Corresponding author: Serban Negoita, MD, DrPH, Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive, Rockville, MD 20850, United States (serban.negoita@nih.gov)

Abstract

Childhood cancers are a heterogeneous group of rare diseases, accounting for less than 2% of all cancers diagnosed worldwide. Most countries, therefore, do not have enough cases to provide robust information on epidemiology, treatment, and late effects, especially for rarer types of cancer. Thus, only through a concerted effort to share data internationally will we be able to answer research questions that could not otherwise be answered. With this goal in mind, the US National Cancer Institute and the French National Cancer Institute co-sponsored the Paris Conference for an International Childhood Cancer Data Partnership in November 2023. This meeting convened more than 200 participants from 17 countries to address complex challenges in pediatric cancer research and data sharing. This Commentary delves into some key topics discussed during the Paris Conference and describes pilots that will help move this international effort forward. Main topics presented include: (1) the wide variation in interpreting the European Union's General Data Protection Regulation among Member States; (2) obstacles with transferring personal health data outside of the European Union; (3) standardization and harmonization, including common data models; and (4) novel approaches to data sharing such as federated querying and federated learning. We finally provide a brief description of 3 ongoing pilot projects. The International Childhood Cancer Data Partnership is the first step in developing a process to better support pediatric cancer research internationally through combining data from multiple countries.

Background

Childhood cancers are a heterogeneous group of rare diseases occurring in people younger than 20 years of age. They represent less than 2% of all cancers, with an estimated 275 713 new cases and 105 345 deaths worldwide in 2022.¹ The relatively small number of cases, even in populous regions, can considerably hamper research efforts to understand the etiologic mechanisms underpinning these cancers. High-dose ionizing radiation, chemotherapy, and prenatal exposure to diethylstilbestrol are the only risk factors that have emerged as definitively causal for childhood cancer.^{2,3}

Advances in diagnosing and treating childhood cancers in recent decades have led to significant survival improvements, with more than 80% of children diagnosed in developed countries becoming 5-year survivors.⁴ This good news, however, predominantly applies to the most common cancers in children, such as acute lymphoblastic leukemia, where more is known about

therapeutic response. Progress for rarer childhood cancers, such as choroid plexus carcinoma, has lagged behind, and only a global collaboration that benefits from increased sample sizes and more research consortia could contribute to meaningful advances.⁵ Furthermore, as more children survive their cancers, short- and long-term sequelae of therapies, such as subsequent cancers, are increasingly being observed in clinical practice and epidemiological studies,^{6,7} making it imperative to follow these patients adequately throughout their lifespan.

Since most countries do not have enough childhood cancer cases to achieve the sample size and genotypic diversity necessary for robust analyses on epidemiology, treatment, and late effects, the US National Cancer Institute (NCI) and the French National Cancer Institute (INCa) obtained institutional support from the European Commission to co-sponsor the Paris Conference for an International Childhood Cancer Data Partnership. This meeting, held in Paris in November 2023,

addressed complex challenges in childhood cancer research and data sharing within the framework of the G7 Cancer initiative. This new international collaboration, led by INCa, involves 7 other organizations (including NCI) and is at the forefront of the fight against cancer.⁸ A major priority for G7 Cancer is to formulate an international data strategy for childhood cancer.

The Paris Conference convened more than 200 participants from many disciplines in pediatric oncology, including clinicians, data scientists, cancer registry and research consortia leaders, and epidemiologists from 17 countries. Participants heard from keynote speakers, patient advocates, cancer survivors, and policymakers, and engaged in fruitful discussions through 4 workshops. A summary report was jointly released on February 15, 2024—International Childhood Cancer Day. It describes key barriers, potential solutions, and next steps that could lead to a large international data resource for childhood cancer research. A major conclusion from the conference is that only through a concerted effort to extend international collaboration and data sharing will we be able to answer research questions that could not otherwise be answered. The Conference also showcased well-established international collaborations focused on advancing childhood cancer research. A notable example is the University of Chicago's Pediatric Cancer Data Commons (PCDC),⁹ the largest single unified platform for childhood cancer clinical trials research in the world.

This Commentary delves into some key topics discussed during these workshops and describes pilots that will help move this effort forward. [Table S1](#) provides the definitions of all the acronyms used in the text.

Legislation

Working Group members agreed that considerable variation exists in interpreting the European Union's General Data Protection Regulation (GDPR)¹⁰ with respect to how consent, public interest, and anonymized data should be defined and applied in secondary analysis of health data for research. The precise meaning of these concepts in the cancer research context and the criteria for correctly interpreting them should be clear and consistent across member states. This consensus aligns with the European Society for Paediatric Oncology (SIOPE) recommendations.¹¹ Many other countries' data privacy laws bear similarities to GDPR or were inspired by its provisions. Examples include Brazil's General Data Protection Law,¹² Japan's Protection of Personal Information Act,¹³ and the United Kingdom's Data Protection Act (UK GDPR).¹⁴ In this Commentary, we focus on GDPR because it is widely considered a reference legal instrument.

Nordic countries interpret the GDPR provisions more conservatively and strictly, and national laws have also made it harder for them to provide patient-level data for international studies.¹⁵ These countries have recently tended to prefer a federated approach for access to such data.¹⁶

In the United States, there were similar challenges with interpreting the privacy provisions of the Health Insurance Portability and Accountability Act (HIPAA).¹⁷ To prevent cancer data reporting and cancer research from being halted due to misinterpretation, the North American Association of Central Cancer Registries (NAACCR) and HIPAA experts developed a summary that provided clear, concise answers to questions about the HIPAA privacy rules in relation to cancer registries and data sharing.¹⁸ This document contributed to a harmonized interpretation of HIPAA across all North American central cancer registries.

The wide variation in interpreting GDPR also applies to informed consent. Article 89 GDPR states that "processing personal data is generally prohibited, unless it is expressly allowed by law, or the data subject has consented to the processing." Cancer registries are usually mandated by law to collect personal health data for the purpose of cancer prevention and control as well as for scientific research conducted in the public interest, so they operate under a consent waiver.

Cancer registration systems requiring informed consent underestimate the true population burden of cancer.¹⁹ In Germany, for example, laws passed in the mid-1980s requiring informed consent from Hamburg and Saarland residents diagnosed with cancer resulted in unacceptably low completeness, with neither region able to collect more than 70% of cancer cases.²⁰ The Working Group discussed the role of individual patient consent in research that is conducted outside of the registry's auspices but still for the common good, such as the secondary use of personal health data aimed at building data resources that will benefit other patients in the future. It may be argued that patients who benefit from a healthcare system have a responsibility to contribute data for the common good without needing their consent.²¹

The Working Group also discussed the concept of broad consent for genomic research and how it is being interpreted differently across countries and institutions, hindering secondary use of clinically acquired data. In brief, broad consent is the act of gaining research participants' consent for multiple potential future research purposes without the obligation to recontact them to request permissions for each new project.²² It is not a waiver but a more flexible alternative to study-specific consent, which requires the research subject to provide consent only to specifically elaborated research projects. For studies requiring approval from an independent ethics committee, also known as institutional review board (IRB) in the United States, researchers usually have 3 options for obtaining consent²³: obtain a waiver of consent, obtain a study-specific consent, or use the broad consent option. The latter provides the most utility to the research process while protecting the subject's autonomous wish to participate in studies that are conducted for the common good. Although GDPR does not specifically include the concept of broad consent, Recital 33 (*Consent to Certain Areas of Scientific Research*) has usually been interpreted as the provision supporting it.²⁴ The European Data Protection Board, however, has adopted a more strict interpretation of Recital 33.²⁵

The Working Group recommended that ethical committees seek a common understanding of broad consent and be empowered to use it as an alternative to study-specific consent whenever the research contributes to the common good and provided that personal data are processed in a lawful, fair, and transparent way, as required by Article 5 GDPR. To help meet these conditions, the research team or consortium should aim to implement an electronic system to support the informed consent process. Compared to paper consent, electronic consent motivates participation, facilitates enrollment, helps keep track of who opts in and who opts out, creates transparency, and increases study knowledge and interactivity.²⁶

With respect to GDPR, anonymized data (ie, data that have been processed to remove personally identifiable information to maintain anonymity) is not subject to the same restrictions placed on the processing of personal data, but there is a high threshold for rendering the data anonymous (Recital 26 GDPR), despite ongoing debate regarding whether data can be anonymized completely.²⁷ In addition to anonymized data, GDPR

provisions include identified data and pseudonymized data. In pseudonymized data, direct identifiers are replaced with a placeholder value (pseudonym) that does not directly identify the patient but for which a crosswalk to the patient identifier exists with the data owner (Recital 28 GDPR).

In the United States, the HIPAA standard is de-identification, not anonymization. These 2 terms, although often used interchangeably, are not the same.²⁸ Both techniques, however, require that a dataset be stripped of identifiable information. HIPAA provides for a research identification code to be assigned, allowing de-identified health information to be linked back to the identity of the patient to which it corresponds, as long as that crosswalk is not made available to the recipient of the dataset.²⁹ This dataset would be de-identified under HIPAA and pseudonymized under GDPR. De-identification can be achieved using either of 2 methods²⁹: safe harbor (removal of all 18 HIPAA-defined identifiers) or expert determination. An example of de-identified data with a link back to the patient's identity can be found in the PCDC resource in which some data contributors send an honest-broker identifier which allows them to send updates to the data.⁹ Only the data contributor has access to the crosswalk of the de-identified data to the individual patient. An example of anonymized data that is completely stripped of any identifiers or pseudonyms and is publicly available for download is the International Agency for Research on Cancer's CISPlus datasets.³⁰ These methods of data transformation can reduce the risk of re-identification, but they cannot remove it entirely. The classic method of database reconstruction can infer individual-level responses from small counts in tabular data to re-identify individuals³¹ and novel technologies such as quantum computing can break currently used cryptography intended to protect data.³²

GDPR attempted to regulate data sharing, but in practice sharing data did not become easier within the European Union (EU) and it became almost impossible for Member States to transfer personal data outside of the EU. Justifications for transferring personal data outside of the EU follow a 3-tier hierarchy, as outlined in Articles 45 (*Adequacy decision*), 46 (*Appropriate safeguards*), and 49 (*Derogations*) of GDPR. The Adequacy decision is the most appropriate but harder to pursue since the European Commission must determine that the non-EU country or organization has an "adequate level" of data protection. Data transfers under an Adequacy decision are assimilated to intra-EU data transmissions, but the decision must be reviewed by the European Commission periodically. The absence of an Adequacy decision covering personal health data in the United States is impeding research progress in many critical fields. For instance, much of the EU contribution to the NCI Cohort Consortium on rare disorders, which includes childhood cancers, was halted after GDPR came into effect, mainly due to European researchers' inability to share data with NCI.³³

Notwithstanding these legal obstacles to international data sharing, the EU and United States continue to strengthen cooperation in the health arena, namely to facilitate health information exchange to support research, innovation, and public health advances, in compliance with the legal frameworks on both sides of the Atlantic.³⁴ This cooperation builds trust at the highest level, increasing the likelihood of sustainable partnerships.³⁵

The Working Group recommends that the European Network of Cancer Registries and other stakeholders like SIOPE collaborate with the European Data Protection Board to develop crucial guidance for processing personal data in scientific research to better harmonize interpretation and implementation of GDPR

across Member States. This work should account for the rules and provisions of the future European Health Data Space,³⁶ which was designed to facilitate data access for research and innovation across Europe.

Standardization, harmonization, and interoperability

Standardization enables registries and research consortia to achieve high-quality, comparable data, while harmonization sets rules for data handling across the cancer data continuum—starting with how they are collected at the point of care, combined from different sources, and finally processed and released for analysis.³⁷ The International Classification of Diseases for Oncology is a classic example of a fundamental standard that allows registries to give the same name and code to any cancer type occurring anywhere in the world.³⁸ This is the first step in producing comparable measures of cancer. In North America, all entities contributing to cancer surveillance follow NAACCR standards, which ensures data comparability across the continent.³⁹ This standard has been used to rapidly build up secondary research data products by NAACCR and NCI that deliver standardized data to researchers for analysis despite widely varying state and federal laws. Standardization and harmonization greatly enhance interoperability, that is, the ability to share and reuse data across multiple systems without losing semantic, contextual, or structural meaning.⁴⁰ Altogether, these 3 processes optimize health outcomes and stand to benefit patients and society the most.

A critical aspect of data comparability is ensuring consistency in how variables are defined and collected from disparate sources. To this end, it helps to distinguish core data items from more complex, detailed data items which, in principle, only higher-resourced institutions can collect. Well-known projects such as the University College London's International Benchmarking of Childhood Cancer Survival by Stage (BENCHISTA),⁴¹ the London School of Hygiene and Tropical Medicine's worldwide surveillance of trends in cancer survival (CONCORD),⁴² and the International Agency for Research on Cancer's Cancer Risk in Childhood Cancer Survivors (CRICCS)⁴³ have defined both core and more detailed data items in their protocols. Depending on the data item type, a tiered approach could be adopted. For instance, the Toronto staging guidelines adopted a tiered hierarchical approach so that what has been collected with more detail can be collapsed and compared to what has been collected with less detail.⁴⁴ This approach could help account for widely varying types of resources and allow researchers to compare information across registries/countries, regardless of their data collection capacity and without losing any data that have been collected with more detail.

Finally, setting minimum standards and harmonizing the data across cancer surveillance systems (ie, registries) and other organizations/initiatives that collect information on childhood cancers (eg, clinical trials cooperative groups) will be critical for establishing common data models (CDMs). CDMs are used to standardize and facilitate the exchange, pooling, sharing, or storing of data from multiple sources, which can then be used for observational and longitudinal studies.^{45,46}

Several CDMs are available for data harmonization and observational healthcare research. Important examples include the Observational Medical Outcomes Partnership (OMOP), maintained by the Observational Health Data Sciences Informatics (OHDSI) international collaborative,⁴⁷ and the National Patient-Centered

Clinical Research Network (PCORnet), maintained by the Patient-Centered Outcomes Research Institute (PCORI).⁴⁸ The former is being adopted internationally, although some US studies like the National COVID Cohort Collaborative (N3C) have also selected OMOP as the canonical model due to its maturity, documentation, and open-source quality.⁴⁹ Also, through its Oncology Workgroup, OHDSI has developed a specific extension for cancer,⁵⁰ providing a platform for standardizing cancer data, including diagnoses, treatments, and outcomes, to enable the conduct of observational cancer studies and identify patient cohorts in a distributed research network.

Some national (eg, The Netherlands Cancer Registry) and regional (eg, Geneva Cancer Registry) cancer registries in Europe have started to map their data to the OMOP CDM to make the data comparable across different studies/entities using the same CDM. The OMOP CDM has also been adopted in some large-scale, EU-funded initiatives such as IDEA4RC,⁵¹ a collaboration among 25 European institutes to establish a data space for rare cancers. In the United States, there has been an effort to map the common data elements (CDEs) in NCI's cancer Data Standards Registry and Repository (caDSR) to the OMOP CDM. The caDSR CDEs have been adopted to harmonize data submitted by cancer centers and registries to the NCI-led National Childhood Cancer Registry.

The complexity of some of these CDMs may impose significant obstacles for researchers and clinicians, so it may be more accessible to develop ad hoc solutions. The INCa's Interoperability and Data Sharing of Clinical and Biological Data in Oncology (OSIRIS) CDM uses a minimum dataset, with only 60 clinical data items.⁵² Although limited, it is flexible and extensible so that other data items can be added based on consensus between experts and stakeholders. The OHDSI Oncology Workgroup is working to achieve interoperability between the OSIRIS CDM and the OMOP CDM.

Novel approaches to data sharing

Developing large, centralized databases with harmonized data from multiple medical institutions or research studies is an effective way to develop population-based resources for childhood cancer research. Because preserving patient privacy is critical to this process, data owners and processors are particularly attentive to this aspect, so much so that, to our knowledge, no breaches of cancer registry data have occurred. Data privacy and security have been one of the principles of cancer registration since its onset.⁵³

Despite the success of centralized approaches, interest in the data visiting approach has grown, whereby data users query data from the source using the data owner's protected platform. Secure data infrastructures that support data visiting allow the owner to retain physical and operational control over their data by keeping them behind their existing firewalls.⁵⁴ Data visiting can be implemented with privacy enhancing technologies (PETs) and permit training and/or deployment of algorithms without physically transferring sensitive data.

The Working Group discussed 2 increasingly used mechanisms for distributed data sharing under the data visiting paradigm—federated querying and federated learning (FL)—as well as emerging PETs used to improve privacy, including synthetic data generation. Next, we briefly describe each approach, including some pros and cons.

Federated queries

Federated querying (FQ) is a decentralized, distributed approach where a trusted third party has been granted access to the data extracts (ie, queries) from external databases across multiple medical institutions.⁵⁴ Using any number of CDMs agreed on by the partners, the trusted broker authenticates and deploys approved queries across distributed databases to generate a virtual analytical file that is stored in a virtual machine and used for downstream analysis. Access to this file is limited to the virtual environment (web interface), providing a bridge between systems while avoiding the “download and compute locally” approach. Privacy is guaranteed because the retrieval of information is limited to what is specified in the query, with the data remaining in its original location, giving the owner and trusted broker complete control over data use and reuse. Many of the federated query systems in the health space take extra measures to ensure privacy⁵⁵: (1) only releasing the amount of data necessary to answer the question; (2) limiting the use of row level information when it is not necessary for the research question; and (3) masking or not returning information when the query results in small cell counts. PCORnet is an example of FQ done in the context of federated ecosystems.⁴⁸

Federated learning

FL is a decentralized, distributed approach under the machine learning (ML) framework that allows multiple institutions or networks to collaboratively train a model (eg, logistic regression or survival) without moving patient data beyond their systems' firewalls.⁵⁶ This feature minimizes access to or release and transfer of protected health information. The ML process occurs locally at each institution, and only the model parameters (ie, gradients) needed to update a global model are transferred back and forth between the sites and the central server coordinated by the trusted broker. Some systems do not require a central server,⁵⁷ but our discussion is limited to the centralized aggregation paradigm. Once the global model has been trained with heterogeneous data from multiple institutions, it can be used, for example, to improve mortality prediction models.⁵⁸ Some notable studies using FL for cancer research include the Federated Tumor Segmentation (FeTS) initiative and the EU-funded FLORENCE project. FeTS is one of the largest FL studies worldwide. Its aim is to generate an automatic tumor boundary detector for glioblastoma, a rare, highly fatal brain tumor.⁵⁹ FLORENCE is a FL project successfully used to enhance colorectal cancer care in the Nordic countries.⁶⁰

Neither FQ nor FL are devoid of security risks. Privacy leakage can still happen if further security measures are not added to protect raw data before and during the processing stage (input privacy) or data that is shared or released after processing, including trained models and model parameters (output privacy). Extra layers of privacy are usually achieved by stacking additional PETs on top of the federated architecture. Techniques to improve input privacy include homomorphic encryption,⁶¹ which enables computation directly on encrypted data, and multi-party computation,⁶² which allows multiple nodes to collaboratively compute a function over their inputs while keeping those inputs private. Techniques to improve output privacy include differential privacy (DP)⁶³ and synthetic data (discussed below). DP is an umbrella term for techniques that, through the injection of random changes, introduce distortions to the data to try and make it unrecognizable while retaining the statistical properties of the original dataset (this is the concept of “adding noise” to the data).

The federated approaches described above allow each participating site to keep physical and operational control over their data, which helps overcome participation barriers. Such approaches would also save costs and time by removing the requirement to transfer huge volumes of raw data to a trusted third party, who must also integrate, manage, and store those data. Finally, once issues with the privacy/accuracy trade-off have been resolved, federated approaches will benefit from increased sample sizes and number of institutions in the distributed network. This is particularly important when studying rare diseases such as childhood cancer. The downside of most of these emerging techniques is that they achieve more robust privacy protection at the expense of accuracy, leading to a privacy/utility trade-off.⁶⁴ They also lead to higher computational and research costs, limiting their applicability in low-resource settings.

Synthetic data

As mentioned, synthetic data generation is an approach to improve output privacy. It involves creating an artificial dataset that retains the statistical properties of a real-world dataset.⁶⁵ Privacy is protected because the data does not represent real individuals and, therefore, does not contain sensitive data or personal identifiers. Synthetic data is useful for gaining insights into the disease represented in the dataset, conducting experiments, and validating models. However, since it only approximates the original data, it should not be used to answer epidemiological questions or make clinical decisions. For that, researchers need to rely on real data. The challenge with creating a high-fidelity synthetic dataset is that there are interdependencies between the variables in real-world data that need to be protected so the synthetic data retains its utility.⁶⁶ This could be a limitation when creating a synthetic dataset for rare diseases such as childhood cancer, where a proper balance between utility and privacy is harder to achieve due to small counts. Building a dataset to answer a specific research question or conduct a specific analysis could help overcome this limitation. When there is a concern that the synthetic data could be correlated with external data to infer sensitive information (a likely scenario when studying rare diseases), additional privacy techniques such as DP may be needed to add an extra layer of protection. Examples of synthetic data include one of the 3 data types available in the N3C study⁴⁹ and Simulacrum,⁶⁷ a dataset that imitates some of the data collected and curated by NHS England's National Disease Registration Service. Synthetic data are expected to fall under the new data governance requirements in the upcoming EU's Artificial Intelligence Act.⁶⁸

Steps forward

As an outcome of the Paris Conference, the NCI and INCa have implemented several workstreams and are working with partners to build and develop pilot projects. The projects will provide use cases to assess what types of data sharing and research are feasible for sharing childhood cancer data internationally. The use cases are generally based on examples of successful, smaller-scale projects. The overarching goals of these pilots are to: (1) provide use cases based on successful prior work that demonstrates the scalability of methods; (2) provide proof of concept to demonstrate data sharing feasibility; (3) demonstrate the ability to enable data access by a broader external research community; and (4) address research questions that could not be

answered without international data sharing. Below we describe 3 of the ongoing pilots.

One pilot involves sharing registry data between the United States, Canada, and France, including measures of rurality and socioeconomic status (SES). This pilot aims to: (1) demonstrate that international data sharing is feasible; (2) assess the differential impact of SES and rurality on cancer outcomes across countries; and (3) make the data accessible to outside researchers through Statistics Canada's Data Analytics Services, a cloud-based, customizable, and highly secure data analytics platform.⁶⁹ Low SES impacts survival outcomes in children with cancer,⁶⁹ but measuring SES is not standardized across countries.

The second pilot is a collaboration between the United States and the United Kingdom that uses PETs to facilitate research on childhood cancers. Using some very rare pediatric cancers as a use case (eg, hepatocellular carcinoma), this pilot aims to: (1) enable cross-border pediatric oncology research; (2) demonstrate that data sharing is possible without exchanging raw data; (3) identify the best trade-off between privacy and accuracy by applying varying levels of differential privacy guarantees; and (4) prove the use of the most appropriate working technology in a real-world application that is economically viable and works as an enabler for research. PETs like FL and FQ can be an effective solution for securely accessing more diverse data while protecting patients' privacy and complying with legal requirements.⁷⁰

The third pilot involves expanding ExtractEHR beyond the United States to work with European data registries. ExtractEHR is an R software package that uses an application programming interface to extract data from electronic health record (EHR) systems, including demographics, clinical notes, laboratory results, medications, pathology and radiology reports, procedures, and adverse events with a laboratory-defining component.⁷¹ This tool has been implemented across multiple US institutions and is being expanded to registries within the NCI's Surveillance, Epidemiology, and End Results (SEER) Program. This pilot aims to implement ExtractEHR in a few European childhood cancer facilities to rapidly identify and extract rich, longitudinal clinical data for cancer surveillance, cohort studies, or clinical trial assessment/enrollment.

Conclusion

In this Commentary, we described the creation of a novel international partnership to facilitate research on childhood cancers. We discussed some technical and legal challenges to international sharing of health data in the context of childhood cancer research, agreed on the need to an improved interpretation of GDPR, described novel methods and technologies, and highlighted ongoing activities promoting international data sharing. A major conclusion from the Paris Conference for an International Childhood Cancer Data Partnership is that while technology is not the main barrier to sharing data internationally, possible methods and tools need to mature to reach a good balance between accuracy (closely related to data utility) and privacy to help drive progress in this space.

The benefits to having centralized data are significant. First and foremost, the centralized approach passed the test of time long ago, with cancer registries and epidemiologists presenting an impeccable track record in maintaining the security of personal data for informative and policy-relevant research over more than 60 years. With such a perfect track record, one may question whether the described novel approaches to data sharing are a defensible way to generate reliable information on data

quality and survival, whether they are an efficient use of analytical resources, or whether they protect the autonomy of data subjects or the security of the data any better than traditional, centralized approaches. Secondly, centralized data allows one to perform standardized quality control of the data and set data benchmarks at a centralized, more detailed, case level to improve data quality control as the collected data items evolve.⁷² Finally, centralized data are needed to periodically test the accuracy of novel approaches like FQ and FL so that the queries and models do not diverge from the baseline.

With the active support and participation of the G7 Cancer stakeholders, including NCI and INCa, as well as the research community, this unique opportunity to leverage childhood cancer research internationally has the potential to address questions requiring large data resources and consortia and achieve scientific benefits that could not otherwise be achieved. Examples of the latter include the development of international risk stratification standards permitting joint clinical trials or comparisons between trials and precise estimation of genomic subtypes' prevalence across ethnicities, countries, and global regions.

Many obstacles and challenges lie ahead, but now is the time to act.

Acknowledgments

The authors would like to thank the discussants who presented their work at the conference, which greatly enriched the debate with participants. We also thank the two anonymous reviewers for their constructive suggestions. We also thank Heidi Hanson (Oak Ridge National Laboratory), Suzi Birz (University of Chicago's Data for the Common Good), and John Smith (UK's Department for Science, Innovation & Technology) for their valuable feedback. We also thank Susan Scott (National Cancer Institute) for assistance with manuscript editing.

Author contributions

Gonçalo Forjaz, DVM, MSc (Conceptualization; Writing—original draft; Writing—review & editing), Betsy Kohler, MPH (Writing—review & editing), Michel P. Coleman, MD (Writing—review & editing), Eva Steliarova-Foucher, PhD (Writing—review & editing), Serban Negoita, MD, DrPH (Writing—review & editing), Jaime M. Guidry Auvil, PhD (Writing—review & editing), Fernanda Silva Michels, MSc, PhD (Writing—review & editing), Johanna Goderre, MPH (Writing—review & editing), Charles Wiggins, PhD (Writing—review & editing), Eric B. Durbin, DrPH, MSc (Writing—review & editing), Gijs Geleijnse, MSc, PhD (Writing—review & editing), Marie-Charlotte Henrion, BA, MSc (Writing—review & editing), Candice Altmayer, MSc (Writing—review & editing), Thomas Dubois, PhD (Writing—review & editing), and Lynne Penberthy, MD, MPH (Supervision; Writing—review & editing).

Supplementary material

Supplementary material is available at JNCI: Journal of the National Cancer Institute online.

Funding

G.F. and C.W. were supported, respectively, by Contracts HHSN261201600004B and HHSN261201800014I (Task Order

HHSN26100001) from the National Cancer Institute; G.G. was supported by the Minderoo Foundation.

Conflicts of interest

The authors declare no conflict of interest. S.N., who is a JNCI Associate Editor and co-author on this paper, was not involved in the editorial review or decision to publish the manuscript.

IARC disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article, and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

Data availability

There are no new data associated with this paper.

References

1. Ferlay J, Ervik M, Lam F, et al. *Global Cancer Observatory: Cancer Today*. Lyon, France: International Agency for Research on Cancer; 2024. <https://gco.iarc.who.int/today>
2. Spector LG, Pankratz N, Marcotte EL. Genetic and nongenetic risk factors for childhood cancer. *Pediatr Clin North Am*. 2015;62:11-25. <https://doi.org/10.1016/j.pcl.2014.09.013>
3. Stiller CA. Epidemiology and genetics of childhood cancer. *Oncogene*. 2004;23:6429-6444. <https://doi.org/10.1038/sj.onc.1207717>
4. Rodriguez-Galindo C, Friedrich P, Alcasabas P, et al. Toward the cure of all children with cancer through collaborative efforts: pediatric oncology as a global challenge. *J Clin Oncol*. 2015;33:3065-3073. <https://doi.org/10.1200/jco.2014.60.6376>
5. Adamson PC. Improving the outcome for children with cancer: development of targeted new agents. *CA Cancer J Clin*. 2015;65:212-220. <https://doi.org/10.3322/caac.21273>
6. Ward E, DeSantis C, Robbins A, Kohler B, Jemal A. Childhood and adolescent cancer statistics, 2014. *CA Cancer J Clin*. 2014;64:83-103. <https://doi.org/10.3322/caac.21219>
7. Reulen RC, Winter DL, Lancashire ER, et al. Health-status of adult survivors of childhood cancer: a large-scale population-based study from the British Childhood Cancer Survivor Study. *Int J Cancer*. 2007;121:633-640. <https://doi.org/10.1002/ijc.22658>
8. Senior K. G7 Cancer: the priorities and challenges ahead. *Lancet Oncol*. 2023;24:e240. [https://doi.org/10.1016/s1470-2045\(23\)00236-x](https://doi.org/10.1016/s1470-2045(23)00236-x)
9. 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;5:1034-1043. <https://doi.org/10.1200/cci.21.00075>
10. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016. <https://eurlex.europa.eu/eli/reg/2016/679/oj>.
11. Vassal G, Lazarov D, Rizzari C, Szczepański T, Ladenstein R, Kearns PR. The impact of the EU General Data Protection

- Regulation on childhood cancer research in Europe. *Lancet Oncol*. 2022;23:974-975. [https://doi.org/10.1016/s1470-2045\(22\)00287-x](https://doi.org/10.1016/s1470-2045(22)00287-x)
12. Canedo ED, Calazans ATS, Bandeira IN, Costa PHT, Masson ETS. Guidelines adopted by agile teams in privacy requirements elicitation after the Brazilian general data protection law (LGPD) implementation. *Requir Eng*. 2022;27:545-567. <https://doi.org/10.1007/s00766-022-00391-7>
 13. Yamamoto N, Kawashima M, Fujita T, Suzuki M, Kato K. How should the legal framework for the protection of human genomic data be formulated?—implications from the revision processes of the Act on the Protection of Personal Information (PPI Act). *J Hum Genet*. 2015;60:225-226. <https://doi.org/10.1038/jhg.2014.121>
 14. Francis B. General Data Protection Regulation (GDPR) and data protection act 2018: what does this mean for clinicians? *Arch Dis Child Educ Pract Ed*. 2020;105:298-299. <https://doi.org/10.1136/archdischild-2018-316057>
 15. Larønningen S, Skog A, Engholm G, et al. Nordcan.R: a new tool for federated analysis and quality assurance of cancer registry data. *Front Oncol*. 2023;13:1098342. <https://doi.org/10.3389/fonc.2023.1098342>
 16. Gini A, Colombet M, de Paula Silva N, et al.; CRICCS Consortium. A new method of estimating prevalence of childhood cancer survivors (POCCS): example of the 20-year prevalence in The Netherlands. *Int J Epidemiol*. 2023;52:1898-1906. <https://doi.org/10.1093/ije/dyad124>
 17. Krzyzanowski B, Manson SM. Twenty years of the health insurance portability and accountability act safe harbor provision: unsolved challenges and ways forward. *JMIR Med Inform*. 2022;10:e37756. <https://doi.org/10.2196/37756>
 18. NAACCR. Frequently Asked Questions and Answers About Cancer Reporting and the HIPAA privacy rule. Springfield, IL; 2003. Accessed June 12, 2024. <https://www.naacr.org/wp-content/uploads/2017/01/FAQs-about-HIPAA-and-Cancer-Registry.pdf>
 19. Illman J. Cancer registries: should informed consent be required? *J Natl Cancer Inst*. 2002;94:1269-1270. <https://doi.org/10.1093/jnci/94.17.1269>
 20. Dudeck J. Informed consent for cancer registration. *Lancet Oncol*. 2001;2:8-9. [https://doi.org/10.1016/s1470-2045\(00\)00185-6](https://doi.org/10.1016/s1470-2045(00)00185-6)
 21. Doll R, Peto R. Rights involve responsibilities for patients. *BMJ*. 2001;322:730.
 22. Hallinan D. Broad consent under the GDPR: an optimistic perspective on a bright future. *Life Sci Soc Policy*. 2020;16:1. <https://doi.org/10.1186/s40504-019-0096-3>
 23. Maloy JW, Bass PF 3rd. Understanding broad consent. *Ochsner J*. 2020;20:81-86. <https://doi.org/10.31486/toj.19.0088>
 24. Rumbold JMM, Pierscionek B. The effect of the general data protection regulation on medical research. Viewpoint. *J Med Internet Res*. 2017;19:e47. <https://doi.org/10.2196/jmir.7108>
 25. European Data Protection Board. Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research. Adopted February 2, 2021. Accessed June 12, 2024. https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_replyec_questionnaire_research_final.pdf
 26. Guarino J, Parvanova I, Finkelstein J. Characteristics of electronic informed consent platforms for consenting patients to research studies: a scoping review. *Stud Health Technol Inform*. 2022;290:777-781. <https://doi.org/10.3233/shti220184>
 27. Groos D, van Veen E. Anonymised data and the rule of law. *Eur Data Protect Law Rev (EDPL)*. 2020;6:498-508. <https://doi.org/10.21552/edpl/2020/4/6>
 28. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *J Med Internet Res*. 2019;21:e13484. <https://doi.org/10.2196/13484>
 29. Burckhardt P, Padman R. deidentify. *AMIA Annu Symp Proc*. 2018;2017:485-494.
 30. Ferlay J, Colombet M, Bray F. Cancer incidence in five continents. *CI5plus: IARC CancerBase No. 9*. International Agency for Research on Cancer. Accessed June 12, 2024. <https://ci5.iarc.who.int>
 31. Ruggles S, Van Riper D. The role of chance in the census bureau database reconstruction experiment. *Popul Res Policy Rev*. 2022;41:781-788. <https://doi.org/10.1007/s11113-021-09674-3>
 32. Zerdick T, Olejnik L, Riemann R. *TechDispatch—Quantum Computing and Cryptography*, Issue 2. European Data Protection Supervisor; 2020.
 33. The European Academies Science Advisory Council (EASAC), the Federation of European Academies of Medicine (FEAM) & the European Federation of Academies of Sciences and Humanities (ALLEA). *International Sharing of Personal Health Data for Research*. Halle-Brussels-Berlin: EASAC/FEAM/ALLEA; 2021. <https://doi.org/10.26356/IHDT>
 34. European Commission. *EU-US Joint Statement of the Trade and Technology Council*. Brussels: European Commission; 2022. Accessed June 12, 2024. https://ec.europa.eu/commission/presscorner/detail/en/statement_22_7516
 35. Kerasidou A. The role of trust in global health research collaborations. *Bioethics*. 2019;33:495-501. <https://doi.org/10.1111/bioe.12536>
 36. Horgan D, Hajdich M, Vrana M, et al. European health data space—an opportunity now to grasp the future of data-driven healthcare. *Healthcare (Basel)*. 2022;10:1629. <https://doi.org/10.3390/healthcare10091629>
 37. Osterman TJ, Terry M, Miller RS. Improving cancer data interoperability: the promise of the minimal common oncology data elements (mCODE) initiative. *JCO Clin Cancer Inform*. 2020;4:993-1001. <https://doi.org/10.1200/cci.20.00059>
 38. Fritz A, Percy C, Jack A, et al. *International Classification of Diseases for Oncology: ICD-O*. 3rd ed. World Health Organization; 2000.
 39. Thornton ML, ed. *Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Version 24*, 25th ed. North American Association of Central Cancer Registries; 2023 (revised Jul 2023, Aug 2023, Oct 2023, Feb 2024, Mar 2024). <https://apps.naacr.org/data-dictionary/data-dictionary/version=24/chapter-view/>
 40. Goel AK, Campbell WS, Moldwin R. Structured data capture for oncology. *JCO Clin Cancer Inform*. 2021;5:194-201. <https://doi.org/10.1200/cci.20.00103>
 41. Botta L, Gatta G, Didonè F, Lopez Cortes A, Pritchard-Jones K; BENCHISTA Project Working Group. International benchmarking of childhood cancer survival by stage at diagnosis: the BENCHISTA project protocol. *PLoS One*. 2022;17:e0276997. <https://doi.org/10.1371/journal.pone.0276997>
 42. Allemani C, Matsuda T, Di Carlo V, et al.; CONCORD Working Group. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*. 2018;391:1023-1075. [https://doi.org/10.1016/s0140-6736\(17\)33326-3](https://doi.org/10.1016/s0140-6736(17)33326-3)
 43. de Paula Silva N, Gini A, Dolya A, et al.; CRICCS consortium. Prevalence of childhood cancer survivors in Europe: a scoping review. *EJC Paediatr Oncol*. 2024;3:None.
 44. Gupta S, Aitken JF, Bartels U, et al. Paediatric cancer stage in population-based cancer registries: the Toronto consensus

- principles and guidelines. *Lancet Oncol*. 2016;17:e163-e172. [https://doi.org/10.1016/s1470-2045\(15\)00539-2](https://doi.org/10.1016/s1470-2045(15)00539-2)
45. Ryu B, Yoon E, Kim S, et al. Transformation of pathology reports into the common data model with oncology module: use case for colon cancer. *J Med Internet Res*. 2020;22:e18526. <https://doi.org/10.2196/18526>
 46. Frid S, Bracons Cucó G, Gil Rojas J, et al. Evaluation of OMOP CDM, i2b2 and ICGC ARGO for supporting data harmonization in a breast cancer use case of a multicentric European AI project. *J Biomed Inform*. 2023;147:104505. <https://doi.org/10.1016/j.jbi.2023.104505>
 47. Biedermann P, Ong R, Davydov A, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. *BMC Med Res Methodol*. 2021;21:238. <https://doi.org/10.1186/s12874-021-01434-3>
 48. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21:578-582. <https://doi.org/10.1136/amiajnl-2014-002747>
 49. Haendel MA, Chute CG, Bennett TD, et al.; N3C Consortium. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021;28:427-443. <https://doi.org/10.1093/jamia/ocaa196>
 50. Belenkaya R, Gurley MJ, Golozar A, et al. Extending the OMOP common data model and standardized vocabularies to support observational cancer research. *JCO Clin Cancer Inform*. 2021;5:12-20. <https://doi.org/10.1200/cci.20.00079>
 51. European Commission. IDEA4RC Project Factsheet. Accessed June 12, 2024. <https://cordis.europa.eu/project/id/101057048>
 52. Guérin J, Laizet Y, Le Texier V, et al. OSIRIS: a minimum data set for data sharing and interoperability in oncology. *JCO Clin Cancer Inform*. 2021;5:256-265. <https://doi.org/10.1200/cci.20.00094>
 53. Muir CS, Démaret E. Cancer registration: legal aspects and confidentiality. In: Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, eds. *Cancer Registration: Principles and Methods*. International Agency for Research on Cancer; 1991:199-207.
 54. Platt R, Brown JS, Robb M, et al. The FDA sentinel initiative—an evolving national resource. *N Engl J Med*. 2018;379:2091-2093. <https://doi.org/10.1056/NEJMp1809643>
 55. Gisslander K, Rutherford M, Aslett L, et al.; FAIRVASC consortium. Data quality and patient characteristics in European ANCA-associated vasculitis registries: data retrieval by federated querying. *Ann Rheum Dis*. 2024;83:112-120. <https://doi.org/10.1136/ard-2023-224571>
 56. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3:119. <https://doi.org/10.1038/s41746-020-00323-1>
 57. Yuan L, Sun L, Yu P, Wang Z. Decentralized federated learning: a survey and perspective. *IEEE Internet Things J*. 2023;11:34617-34638.
 58. Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med*. 2021;27:1735-1743. <https://doi.org/10.1038/s41591-021-01506-3>
 59. Pati S, Baid U, Edwards B, et al. Federated learning enables big data for rare cancer boundary detection. *Nat Commun*. 2022;13:7346. <https://doi.org/10.1038/s41467-022-33407-5>
 60. FLORENCE: About the Project. 2024. Accessed June 12, 2024. <https://florence.forskning.eu/en/>
 61. Blatt M, Gusev A, Polyakov Y, Goldwasser S. Secure large-scale genome-wide association studies using homomorphic encryption. *Proc Natl Acad Sci USA*. 2020;117:11608-11613. <https://doi.org/10.1073/pnas.1918257117>
 62. Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. *Nat Biotechnol*. 2018;36:547-551. <https://doi.org/10.1038/nbt.4108>
 63. Qi T, Wu F, Wu C, He L, Huang Y, Xie X. Differentially private knowledge transfer for federated learning. *Nat Commun*. 2023;14:3785. <https://doi.org/10.1038/s41467-023-38794-x>
 64. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nat Commun*. 2021;12:5910. <https://doi.org/10.1038/s41467-021-25972-y>
 65. Walonoski J, Kramer M, Nichols J, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc*. 2018;25:230-238. <https://doi.org/10.1093/jamia/ocx079>
 66. Vallevik VB, Babic A, Marshall SE, et al. Can I trust my fake data—a comprehensive quality assessment framework for synthetic tabular data in healthcare. *Int J Med Inform*. 2024;185:105413. <https://doi.org/10.1016/j.ijmedinf.2024.105413>
 67. Horvat P, Gray CM, Lambova A, et al. Comparing findings from a friends of cancer research exploratory analysis of real-world end points with the cancer analysis system in England. *JCO Clin Cancer Inform*. 2021;5:1155-1168. <https://doi.org/10.1200/cci.21.00013>
 68. European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts 2021 [COM(2021) 206 final]. Accessed June 12, 2024. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>
 69. Gupta S, Wilejto M, Pole JD, Guttmann A, Sung L. Low socioeconomic status is associated with worse survival in children with cancer: a systematic review. *PLoS One*. 2014;9:e89482. <https://doi.org/10.1371/journal.pone.0089482>
 70. Jordan S, Fontaine C, Hendricks-Sturup R. Selecting privacy-enhancing technologies for managing health data use. *Front Public Health*. 2022;10:814163. <https://doi.org/10.3389/fpubh.2022.814163>
 71. Miller TP, Getz KD, Li Y, et al. Rates of laboratory adverse events by course in paediatric leukaemia ascertained with automated electronic health record extraction: a retrospective cohort study from the Children's Oncology Group. *Lancet Haematol*. 2022;9:e678-e688. [https://doi.org/10.1016/s2352-3026\(22\)00168-5](https://doi.org/10.1016/s2352-3026(22)00168-5)
 72. Pfaff ER, Girvin AT, Gabriel DL, et al.; N3C Consortium. Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. *J Am Med Inform Assoc*. 2022;29:609-618. <https://doi.org/10.1093/jamia/ocab217>