original reports abstract

Development of a Data Model and Data Commons for Germ Cell Tumors

Bo Ci, PhD[1]; Donghan M. Yang, PhD[1]; Mark Krailo, PhD[2,3]; Caihong Xia, PhD[3]; Bo Yao, PhD[1]; Danni Luo, MS[1]; Qinbo Zhou, PhD[1]; Guanghua Xiao, PhD[1,4]; Lin Xu, PhD[1]; Stephen X. Skapek, MD[5,6]; Matthew J. Murray, MB BChir, PhD[7]; James F. Amatruda, MD, PhD[2,8]; Lindsay Klosterkemper, MSc[9]; Furqan Shaikh, MSc, MD[10]; Cecile Faure-Conter, MD[11]; Brice Fresneau, MD, PhD[12]; Samuel L. Volchenboum, MD, PhD[13]; Sara Stoneham, MD[14]; Luiz Fernando Lopes, MD, PhD[15]; James Nicholson, MB BChir, MA, DM[16]; A. Lindsay Frazier, MD, ScM9[9]; and Yang Xie, MD, PhD[1,4,6]

Germ cell tumors (GCTs) are considered a rare disease but are the most common solid tumors in adolescents and young adults, accounting for 15% of all malignancies in this age group. The rarity of GCTs in some groups, particularly children, has impeded progress in treatment and biologic understanding. The most effective GCT research will result from the interrogation of data sets from historical and prospective trials across institutions. However, inconsistent use of terminology among groups, different sample-labeling rules, and lack of data standards have hampered researchers' efforts in data sharing and across-study validation. To overcome the low interoperability of data and facilitate future clinical trials, we worked with the Malignant Germ Cell International Consortium (MaGIC) and developed a GCT clinical data model as a uniform standard to curate and harmonize GCT data sets. This data model will also be the standard for prospective data collection in future trials. Using the GCT data model, we developed a GCT data commons with data sets from both MaGIC and public domains as an integrated research platform. The commons supports functions, such as data query, management, sharing, visualization, and analysis of the harmonized data, as well as patient cohort discovery. This GCT data commons will facilitate future collaborative research to advance the biologic understanding and treatment of GCTs. Moreover, the framework of the GCT data model and data commons will provide insights for other rare disease research communities into developing similar collaborative research platforms.

JCO Clin Cancer Inform 00:1-10. © 2020 by American Society of Clinical Oncology

INTRODUCTION

Germ cell tumors (GCTs) are rare, yet they account for 15% of all malignancies diagnosed during adolescence.[1,2] Although the advent of platinum-based therapy has largely improved the survival rate for patients with GCTs,[3-7] 15% to 20% do not respond.[8] Furthermore, significant long-term effects are associated with current platinum-based chemotherapy treatment. Survivors of GCTs have a significantly elevated risk of developing cardiovascular disease[9] and secondary malignancies.[10] Therefore, there is an urgent need to advance the understanding of the biologic mechanisms of GCTs and develop new therapeutics with better efficacy and fewer adverse effects.

Currently, the clinical research on GCTs has been hampered by their rarity, particularly in younger patients in whom the diversity in site and histologic subtype is also most marked. Among children age 0 to 4 years, the incidence of extracranial GCTs is 7.0 and 5.8 per million for males and females, respectively. The incidence in those age 15 to 19 years is 31 and 25.3 per million for males and females, respectively.[11] The rarity of GCTs makes it impractical for a single institution to run large-scale GCT clinical trials or research projects. Therefore, multi-institutional collaboration is essential for successful research endeavors into GCTs.

Dedicated to facilitating and promoting collaborative projects in GCTs, the Malignant Germ Cell International Consortium (MaGIC) was formed in 2009 by an international group of pediatric and adult oncologists, surgeons, pathologists, epidemiologists, statisticians, bioinformaticians, and basic scientists. Subsequently, MaGIC has initiated several international clinical trials recruiting patients across age groups, countries, and continents. Furthermore, MaGIC is developing a unified data resource for GCT research by leveraging the legacy data sets contributed by its members. Because these legacy data sets were generated independently by different clinical trial organizations, the variables and terminologies vary largely across different data contributors. For instance, Figure 1 shows the discrepancy in the coding of overall histology across three studies: Brazilian clinical trial TCG 99,[12] French trial TGM 95,[13] and merged clinical trials from the United States and United Kingdom.[14] Of 18 apparent valid values for overall histology, only five were shared by all three studies. Even for these five shared values, further confirmation was required to confirm the intended

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on April 29, 2020 and published at ascopubs.org/journal/cci on XXXX, 2020: DOI https://doi.org/10.1200/CCI.20.00025

CONTEXT

Key Objective

To develop a clinical data model for curating and harmonizing germ cell tumors (GCTs) research data sets, and to develop a GCT data commons as an integrated research platform that supports functions, such as data query, management, sharing, visualization, and analysis.

Knowledge Generated

The GCT data model and data commons developed in this work can serve as a standard platform for the integration, management, and prospective generation of GCT data sets. The framework and experience gained here will provide insights for other rare disease research communities in similar endeavors.

Relevance

The rarity of GCTs in children and the lack of interoperability in the existing GCT data sets highlight the great need of a standard for curating and harmonizing GCT clinical research data. This work addresses such need by developing a clinical data model and a data commons where multiple GCT data sets were harmonized and made available for the research community.

meaning for each value across the three studies. For instance, a mixed GCT is a nebulous term that requires further standardization. Therefore, standardization of variable definition and controlled terminology are essential for data sharing and integration across different data contributors and studies.

In recent years, several data commons have been developed as public resources to facilitate biomedical research in a variety of diseases. Most of the existing data commons are focused on cancer, such as the National Cancer Institute (NCI) Genomic Data Commons (GDC), University of California Santa Cruz Xena, cBioPortal, International Cancer Genome Consortium, Catalogue of Somatic Mutations in Cancer, and FireBrowse. Two other data commons are not focused on specific disease areas: European Genome-Phenome Archive and Gene Expression Omnibus (GEO). Although generally not well defined, a data commons usually includes data sets, computing environments (cloud or high-performance computing), software services/tools, and digital objects in compliance with the FAIR standard (Findable, Accessible, Interoperable, and Reusable).[15] In this study, we developed a data model for core clinical GCT variables, which served as a uniform standard to curate and harmonize data sets from MaGIC members. More importantly, this data model could become the standard for prospective data collection in future GCT clinical trials. Building on the data model, we developed a GCT data commons with functionalities like patient cohort discovery or patient query function, data access, management, sharing, visualization, and analysis. The goal was to create an integrated online platform where researchers could find data of interest and perform exploratory analyses. Our data commons will empower data sharing and research collaboration to advance GCT research. In addition, the experience and workflow of developing the GCT data model and commons will benefit similar efforts in other types of rare disease.

METHODS

Design of the GCT Concept Map

As the first step of data model development, we reviewed the clinical episodes that patients with GCTs may experience based on the literature and input from experienced clinicians. Following the concept map for breast cancer (TAUG-BrCa version 1.0) developed by the Clinical Data Interchange Standards Consortium (CDISC), we designed a GCT concept map (Fig 2) for GCT clinical episodes.

Development of the GCT Data Model

The GCT data model was developed through a collaborative effort by MaGIC members based on the GCT concept map. The initial draft of the data model elements, variables, and controlled terminology was first proposed based on the MaGIC data set (N = 1,434, mainly from several clinical trials in the United States and United Kingdom). Several rounds of revision and refinement were performed subsequently with input from MaGIC investigators through teleconferences and international meetings to reach a consensus among the experts in the GCT research community. To ensure the interoperability of the GCT data commons with other data commons and research resources, we defined the treatment-related terminology according to the controlled terminology of CDISC. We further mapped the controlled terminology in the GCT data model to the concept unique identifiers (CUIs) in the NCI Metathesaurus (NCIm) database,[16] which is a biomedical terminology database offering concept mapping to other standards (eg, CDISC and the Systematized Nomenclature of Medicine–Clinical Terms).

Evaluation of the GCT Data Model

The developed data model was used to convert a GCT clinical trial data set from Brazil to test its usability. The Brazilian data set contained pooled data from three clinical trials: TCG91, TCG99, and TCG2008. There were 90

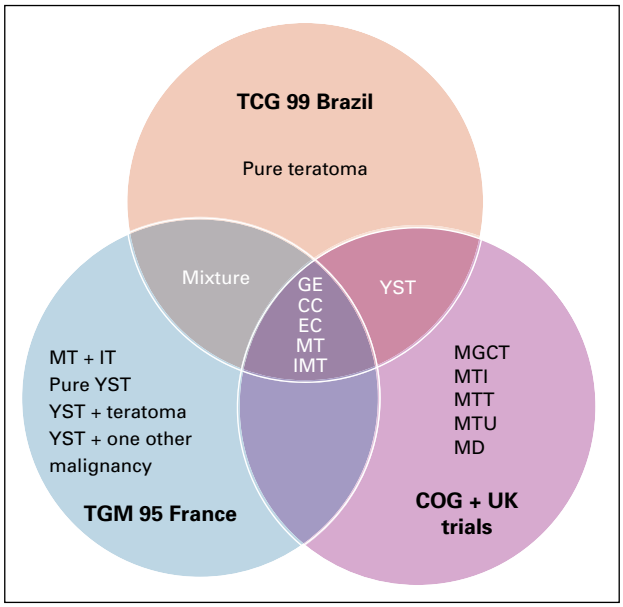


FIG 1. Discrepancy in the valid values for the variable overall histology across three studies. CC, choriocarcinoma; COG, Children's Oncology Group; EC, embryonic carcinoma; GE, germinoma; IMT/IT, immature teratoma; MD, teratoma differentiated; MGCT, mixed germ cell tumor; MT, mature teratoma; MTI, malignant teratoma intermediate; MTT, malignant teratoma trophoblast; MTU, malignant teratoma undifferentiated; YST, yolk sac tumor.

variables after excluding two identifier variables (patient initials and date of birth). We used the number of variables that were successfully mapped to the GCT data model in the Brazilian data set as an evaluation metric.

Construction of the GCT Data Commons

Both MaGIC and the publicly available GCT data sets were imported into the database of the GCT data commons. Public data sets included the Testicular Germ Cell Tumor (TGCT) data set (N = 150) from The Cancer Genome Atlas (TCGA; downloaded from the NCI GDC), the Bagrodia et al[17] data set (N = 180; downloaded from cBioPortal), and the Palmer et al[18] data set (N = 34; downloaded from GEO). All clinical data were curated according to the aforementioned GCT data model. Data format conversion and curation were conducted in the R environment (https://www.r-project.org/). The database and related works were constructed using MySQL (https://www.mysql.com) and PHP (https://www.php.net). The Web portal was developed using JavaScript and Bootstrap (https://getbootstrap.com/docs/3.4/javascript/). The circos plot was constructed using BioCircos.js (http://bioinfo.ibp.ac.cn/biocircos/).

RESULTS

Concept Map of GCT Clinical Episodes

The concept map outlines the general clinical practice workflow for GCT diagnosis, treatment, evaluation, and events, with the red-framed episodes captured in the GCT data model (Fig 2). Patients suspected of having a GCT diagnosis will first go through the diagnostic procedures, which may involve inquiry about demographic information, characterization of primary and metastatic tumor(s) by imaging, and measurements of serum tumor marker levels, including traditional biochemical markers such as alpha fetoprotein (AFP), human chorionic gonadotrophin (HCG), and lactate dehydrogenase (LDH), as well as more novel molecular markers such as microRNA (miRNA) levels. The diagnosis summary will inform and determine the first-line treatment plan. Three types of first-line treatments (surgery, chemotherapy, and radiotherapy) are captured in the data model. After the first-line treatment(s), assessment of treatment response may involve second-look surgery (ie, surgery performed after first-line treatment to determine whether viable tumor remains), radiologic imaging, and measurements of serum tumor marker and miRNA levels. If there is no evidence of disease, the patient will likely be observed via routine GCT monitoring. If the evaluation results are positive for residual cancer, relapse, or a second malignant neoplasm, they will likely trigger a new cycle of disease diagnosis, treatment, and evaluation. Moreover, the first-line and/or second-look surgery will likely generate specimens, which may be used for pathology review and molecular profiling and contribute to the diagnosis summary (Fig 2).

GCT Data Model

Data elements and variables. The GCT data elements refer to functional divisions in the data model, which, according to the clinical episodes, include demographics, disease characteristics, pathology (capturing both central and institutional reviews, if available), serum tumor marker, miRNA level, surgery, chemotherapy, radiotherapy, radiologic response, relapse, second malignant neoplasm, and death or follow-up. The representative variables in each data element are illustrated in Figure 3. Except for demographics, all elements may have $>$ 1 record for each patient. The elements and records can be linked through the data commons patient ID. The element-wise age variables (eg, age at enrollment and age at diagnosis) can be used to rebuild the sequential relationship between episodes. (Note that Fig 3 is an illustrative example of the GCT data model, but not the full version).

Controlled terminology. A controlled terminology (or controlled vocabulary) defines the valid values for a categorical variable, which can be used to standardize the inputs of the categorical variables. Here, we developed the controlled terminology considering both value use frequency and importance based on broad discussions with GCT experts. Different data models can use different sets and combinations of controlled terminologies. Multiple combinations eventually bring challenges to exchanging data between two data models using different terminologies to describe the same variable(s) or concept(s). To facilitate data sharing and communication with other data models/standards, we mapped the valid values in the GCT data model to the CUIs in NCIm (Appendix Fig A1), which permits

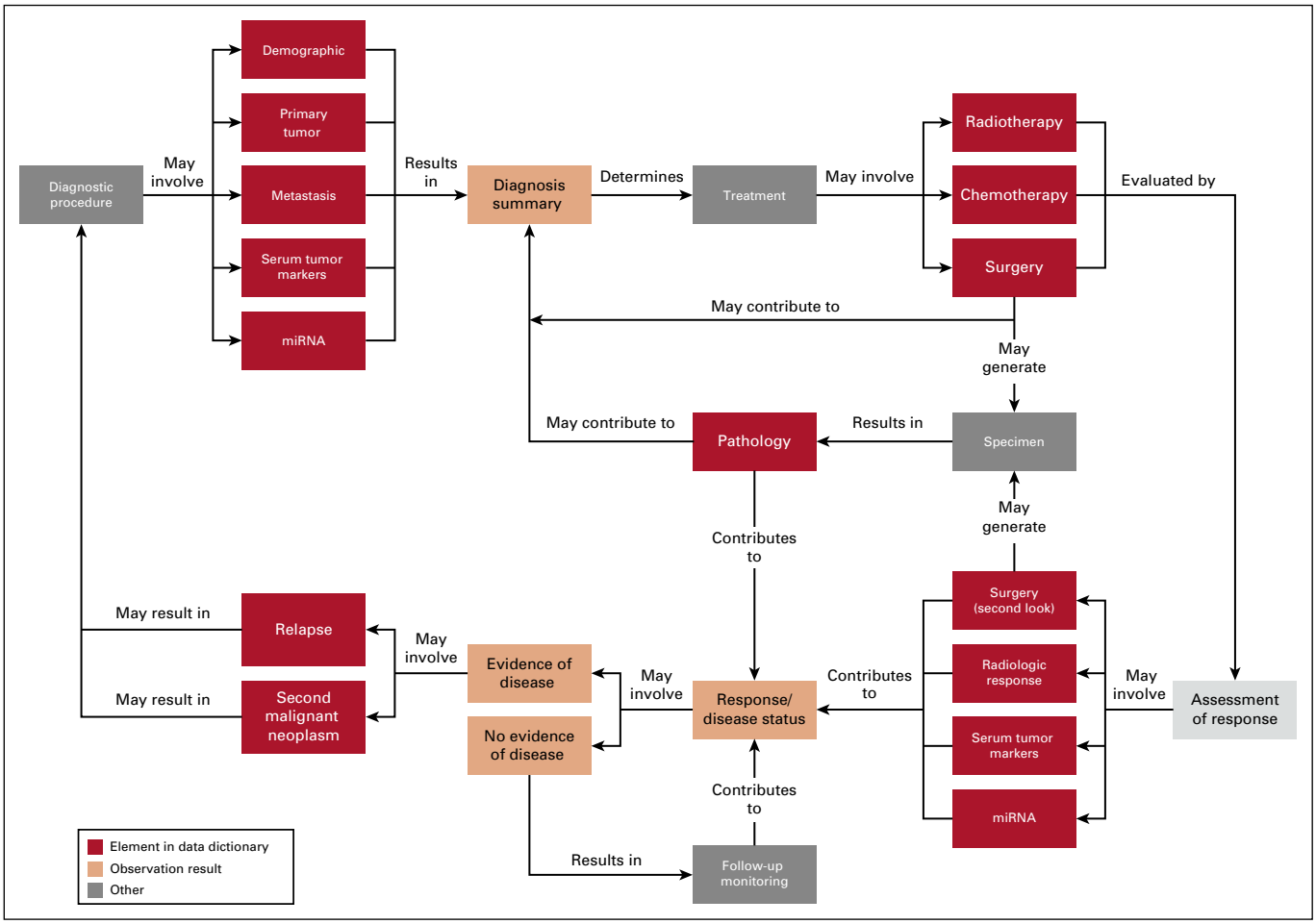


FIG 2. Concept map of germ cell tumor clinical episodes. miRNA, microRNA.

tunneling crosstalk between the GCT data model and other data models/standards.

Evaluation of the GCT Data Model Using a Brazilian Clinical Trial Data Set

To evaluate the usability of the GCT data model, we converted a clinical trial data set from Brazil. The data set had 90 variables, including an anonymous personal identifier (one variable), demographic information (one variable), disease characteristics (15 variable), surgery records (four variables), serum tumor marker levels (44 variables), chemotherapy treatment (seven variables), treatment response (14 variables), and vital status (11 variables). In total, 79 (87.8%) of 90 variables in the Brazilian data set were successfully mapped onto the GCT data model. The large coverage ratio of 87.8% for the Brazilian data set indicates the comprehensiveness of the GCT data model. The variables that were not mapped included two disease characteristic variables, three serum tumor maker variables, and six treatment response variables. The definitions of these 11 variables were not readily available and required further confirmation from the data contributors, which hindered the mapping effort. If the definitions are found to be consistent with the existing ones in the GCT data model, these 11 variables will be mapped as well. Six of the 11 unmapped variables were related to treatment response, which indicates the data model may need further improvement in capturing complex response information.

GCT Data Commons

Harmonized data set. Data on 1,798 patients, 370 samples, and 835 genomic profiles from 14 clinical trials and four programs were incorporated into the GCT data commons. Data types included clinical annotations, single-nucleotide variants (SNVs), copy-number variants (CNVs), mRNA, and microRNA expression.

Cohort discovery. To facilitate the data querying, a cohort discovery module was developed (Fig 4). Users can set the criteria via a filter selection menu (blue frame in Fig 4). The user-selected filters will be automatically displayed in the red frame area in Figure 4 with an unselect option for individual or all filters. The number of available patients and samples will be updated in real time according to the selection criteria (green frame in Fig 4). The summary statistics of the selected subgroup (shown in pie charts) will be

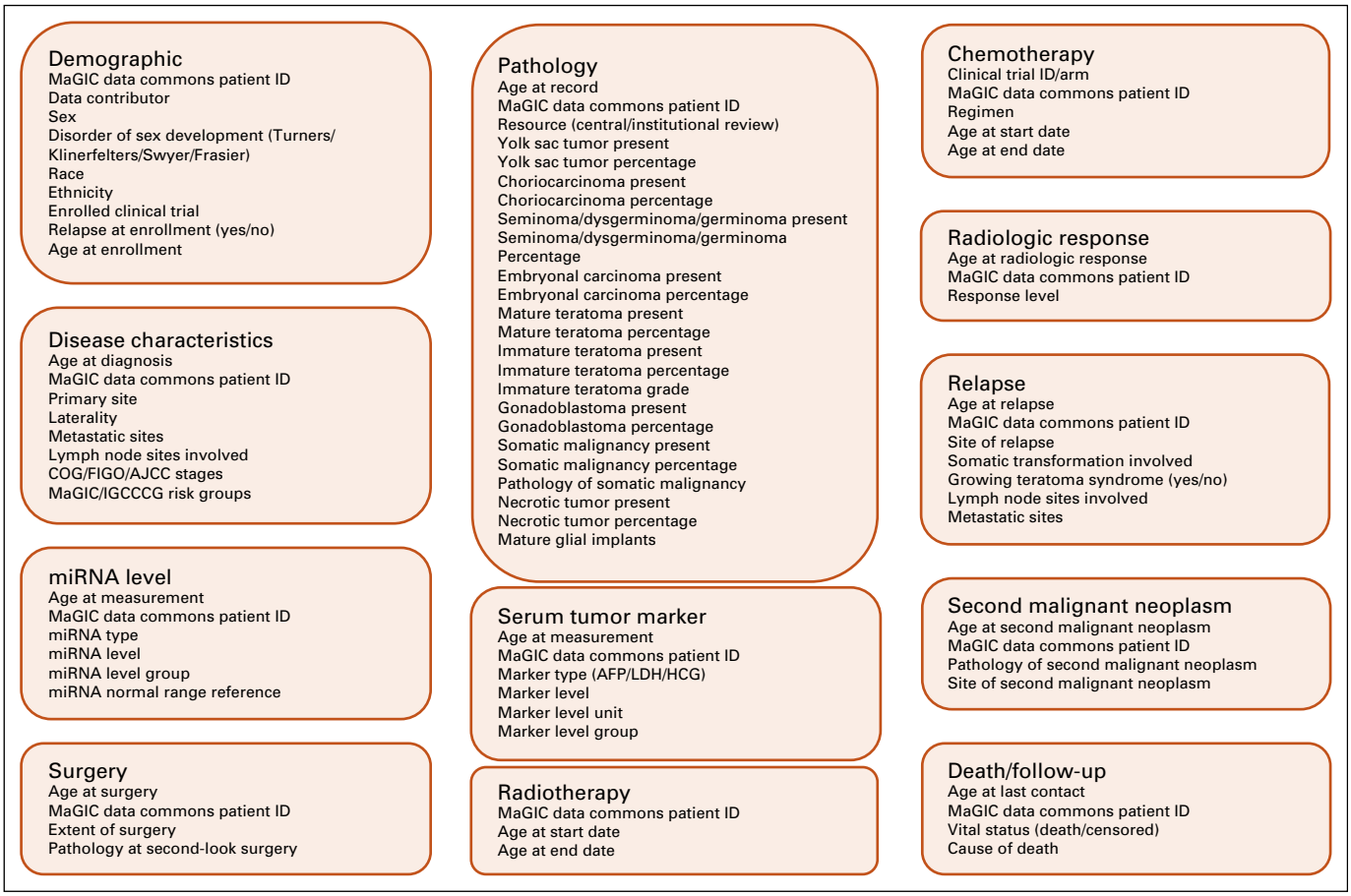


FIG 3. An illustration of data elements and the representative variables in the germ cell tumor data model. AFP, alpha fetoprotein; AJCC, American Joint Committee on Cancer; COG, Children's Oncology Group; FIGO, International Federation of Gynecology and Obstetrics; HCG, human chorionic gonadotrophin; IGCCCG, International Germ Cell Cancer Collaborative Group; LDH, lactate dehydrogenase; MaGIC, Malignant Germ Cell International Consortium; miRNA, microRNA.

updated simultaneously (orange frame in Fig 4). With this module, users may narrow the whole cohort to a specific subgroup of interest. Currently, two types of filters are available: clinical variables (sex, race, age at diagnosis, histology, primary site, relapse, and vital status) and availability of genomic data (SNVs, CNVs, mRNA, and microRNA data) in the data commons.

Visualization modules. To help users gain intuitive views and insights from data, we implemented visualization modules in the GCT data commons. First, the cohort discovery module provides pie charts (orange frame in Fig 4) for the summary of clinical characteristics like sex, age at diagnosis, and histology. Second, the timeline visualization module provides a longitudinal view of detailed clinical information for an individual patient (Fig 5). In the representative patient timeline (Fig 5), the green, blue, and black lines display the time series data of tumor markers LDH, HCG, and AFP, respectively. The colored textboxes label the clinical episodes for the patient with dates on the *x*-axis. When a computer cursor hovers over a textbox, more detailed information related to that clinical episode will appear. For example, a cursor hovering over a tumor diagnosis textbox will show the primary tumor site at diagnosis, and a cursor hovering over a chemotherapy textbox will show the regimen received. The displayed time window can be adjusted by changing the zoom option (top left), entering the date range (top right), or moving the slider (bottom panel; Fig 5). This timeline visualization module enables a user to quickly review the medical history of a given patient together with the changes in associated serum tumor marker levels. Last, the GCT data commons offers a graphical summary of the genomic information for a patient cohort. Figure 6 shows a representative circos plot of the TCGA TGCT patient cohort (N = 156). The circles from inside to outside stand for log-transformed SNV frequency, CNV loss, CNV diploid, CNV gain, mRNA expression, and chromosome location. The circos plot gives users a quick overview of the genomic information in the selected cohort.

Data Security

Security of confidential patient information is paramount. To achieve security, only deidentified data are stored in the

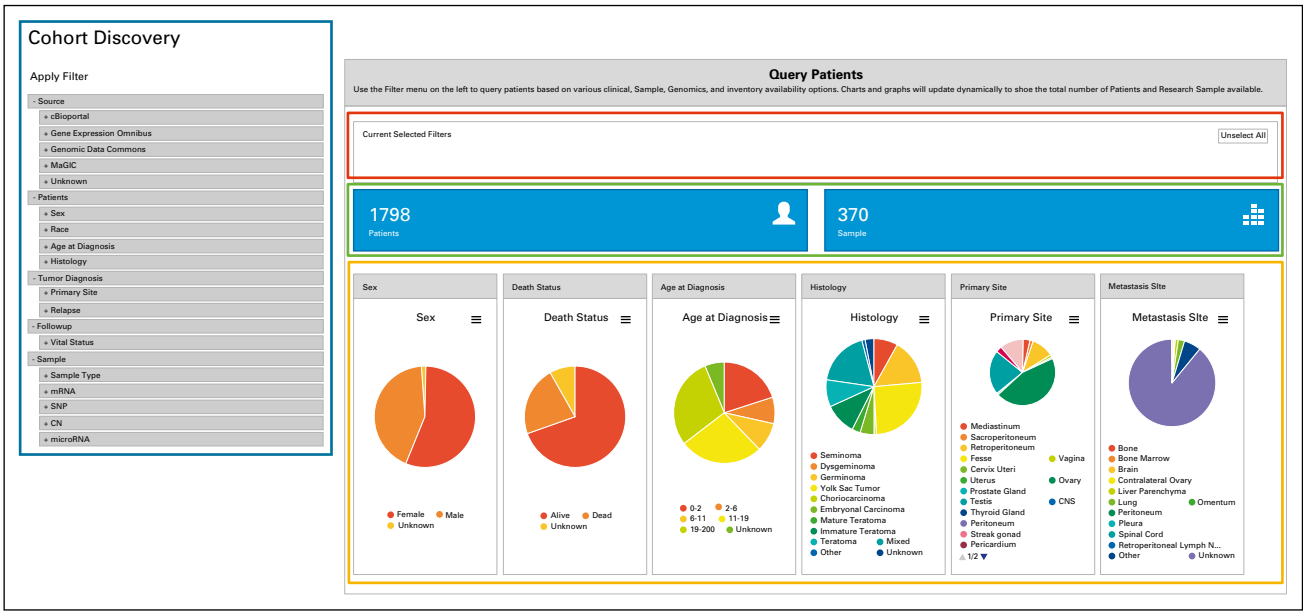


FIG 4. The cohort discovery module of the germ cell tumor data commons. Blue frame, selection filters; red frame, current selected filters; green frame, number of patients and samples in the selected subgroup; orange frame, summary pie charts of clinical variables for the selected subgroup. CN, copy number; MaGIC, Malignant Germ Cell International Consortium; SNP, single-nucleotide polymorphism.

GCT data commons. Second, stringent data storage, server specification, data access, and management protocols are used to protect the data. Only summary statistics, such as patient numbers (Fig 4), and high-level genomic summary information (Fig 6) are available online for the public. Full access to the data requires registration, and access to private data sets requires approval from the MaGIC study committees, according to the standard operating procedures produced by MaGIC. No individual patient data are accessible without registration and appropriate approvals.

Data Model for GCT Specimen

In addition to the main GCT data model, we developed an accompanying data model focusing on specimen information. This specimen data model covers data on the specimen contributor, the anatomic site and laterality, the procedure that generated the specimen (eg, biopsy or surgery) and its corresponding date, the pathologic type (eg, primary solid tumor, metastatic tumor, or solid tissue normal), the preparation method (eg, flash frozen or formalin fixed paraffin embedded), the specimen type (eg, tissue, fluid, cell, or DNA), the amount and concentration (if applicable), and the storage location (Appendix Fig A2). The specimen data model will serve as the joint between clinical data and other data domains when the GCT data commons expands toward specimen-derived data, such as digital pathology images. The information on specimen availability will be shared with users to promote specimen sharing and use across research groups.

DISCUSSION

The concept map of clinical episodes demonstrates the journey that patients with GCTs may experience from first diagnosis to final evaluation (Fig 2). The concept map is essential not only for understanding the general clinical process but also for developing and evaluating the GCT data model. The current concept map was designed to accommodate some specific characteristics of the MaGIC data sets. For example, the measurement of miRNA levels was included for both diagnosis and evaluation procedures, even though it has not been implemented in routine clinical practice. The GCT concept map covers multiline treatments and different types of diagnostic methods as well as evaluations (eg, pathologic review and tumor markers). Although developed for GCTs, it can be adapted to facilitate the development of data models for other diseases.

The GCT data model developed in this study is the first of its kind to our knowledge to focus on clinical data for GCT research, which can also be used as a data exchange standard. It captures typical clinical episodes (eg, diagnostic procedures, treatments, response evaluations, and events) using refined variables and controlled terminology. It took us 3 years, through many rounds of discussion and communication with GCT experts, to reach a consensus on the GCT data model. It is a community effort with essential inputs and detailed suggestions from the GCT research community. An important lesson we learned from this process that may speed up the process in the future is to develop and use a concept map. A good concept map reflects a patient journey by representing the elements and their relationships. Leveraging concept maps in the discussion allows participants to keep the big picture in mind while discussing the finer details of the data elements. Moreover, it could also serve as a blueprint to develop the framework of the data model.

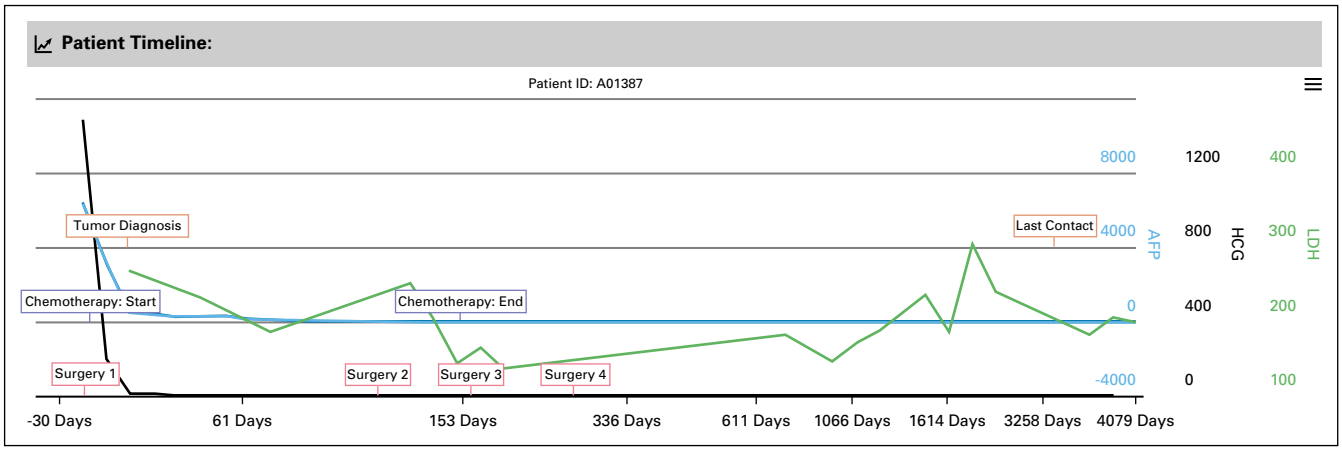


FIG 5. The timeline visualization module of the germ cell tumor data commons; representative timeline for one patient. Green line, timeline of lactate dehydrogenase (LDH) values; black line, timeline of human chorionic gonadotropin (HCG) values; blue line, timeline of alpha fetoprotein (AFP) values. Colored textboxes indicate clinical episodes, such as tumor diagnosis, treatment, and last contact. More detailed information will be shown when a cursor hovers over a textbox.

Implementing the GCT data model will lead to a uniform and standardized data format for storage as well as exchange. This data model has now become the standard for MaGIC members. Investigators outside MaGIC may also adapt the data model in their data collection and curation processes. Broad application of the GCT data model in the research community will greatly improve data interoperability and facilitate data sharing and collaboration. Importantly, this data model may become a standard for prospective data collection in future GCT clinical trials, which will not only save time for new trials but also make the clinical research forms more uninform across different trials.

By design, the GCT data model is a research-orientated development focusing on the clinical experience of patients with GCTs. Data conformed to the GCT data model are readily usable for common data analysis tasks in GCT research. Recent developments in clinical informatics also emphasize building generic, flexible data models or exchange standards for the interoperability of electronic health record (EHR) data. For example, the Informatics for Integrating Biology and the Bedside (i2b2),[19] sponsored by the National Centers for Biomedical Computing, is a generic data model that can handle diverse data types across different diseases and clinical episodes through a star schema. The i2b2 data model offers excellent versatility by storing almost all types of EHR data in a highly denormalized schema. To meaningfully explore, extract, and analyze the data stored in the i2b2 data model, one needs to develop a middle layer of information model[20] that is specific to the question being asked (eg, type of disease or analysis). In another example, the Fast Healthcare Interoperability Resources (FHIR),[21] developed by the Health Level Seven International, is a comprehensive information exchange standard of EHR data aiming to simplify implementation while preserving information integrity. FHIR features comprehensive specifications for storing and transmitting EHR data across EHR operating systems, which offers great flexibility but is by nature not tailored to a specific research area or disease type, like GCTs. A meaningful transfer of data packaged by the FHIR standard (ie, an FHIR resource) still requires the user (typically a health care system or company) to specifically define which information pieces are to be included. In contrast, the GCT data model is designed to provide specific data elements and variables for addressing common data analysis tasks in a specific disease area.

The different emphases of the GCT data model and generic EHR data models provide an excellent opportunity for connecting these two types of informatic efforts together and advancing data interoperability in the GCT research community. The expertise and efforts already infused in the GCT data model can be readily leveraged when connecting with i2b2 or FHIR. For example, the GCT data model can be used as an i2b2 information model, which allows automatic extraction of raw data stored in an i2b2 instance into the research-oriented structure defined by the GCT data model. As a result, institutions that have already implemented the i2b2 infrastructure can easily transfer and merge GCT-related data for research purposes. Similarly, the GCT data model can be represented using one or more resources as defined by FHIR for data transfer purposes. FHIR-based data exchange has been greatly simplified through application programming interfaces. Major EHR operating systems (eg, Epic, Cerner, and MEDITECH) and government agencies (eg, Office of the National Coordinator for Health Information Technology, Centers for Medicare and Medicaid Services, and National Institutes of Health) have adopted the FHIR standards for different uses. Once made compatible with FHIR, continuous transfer of data in the GCT data model can be plugged into existing EHR operating systems and performed across

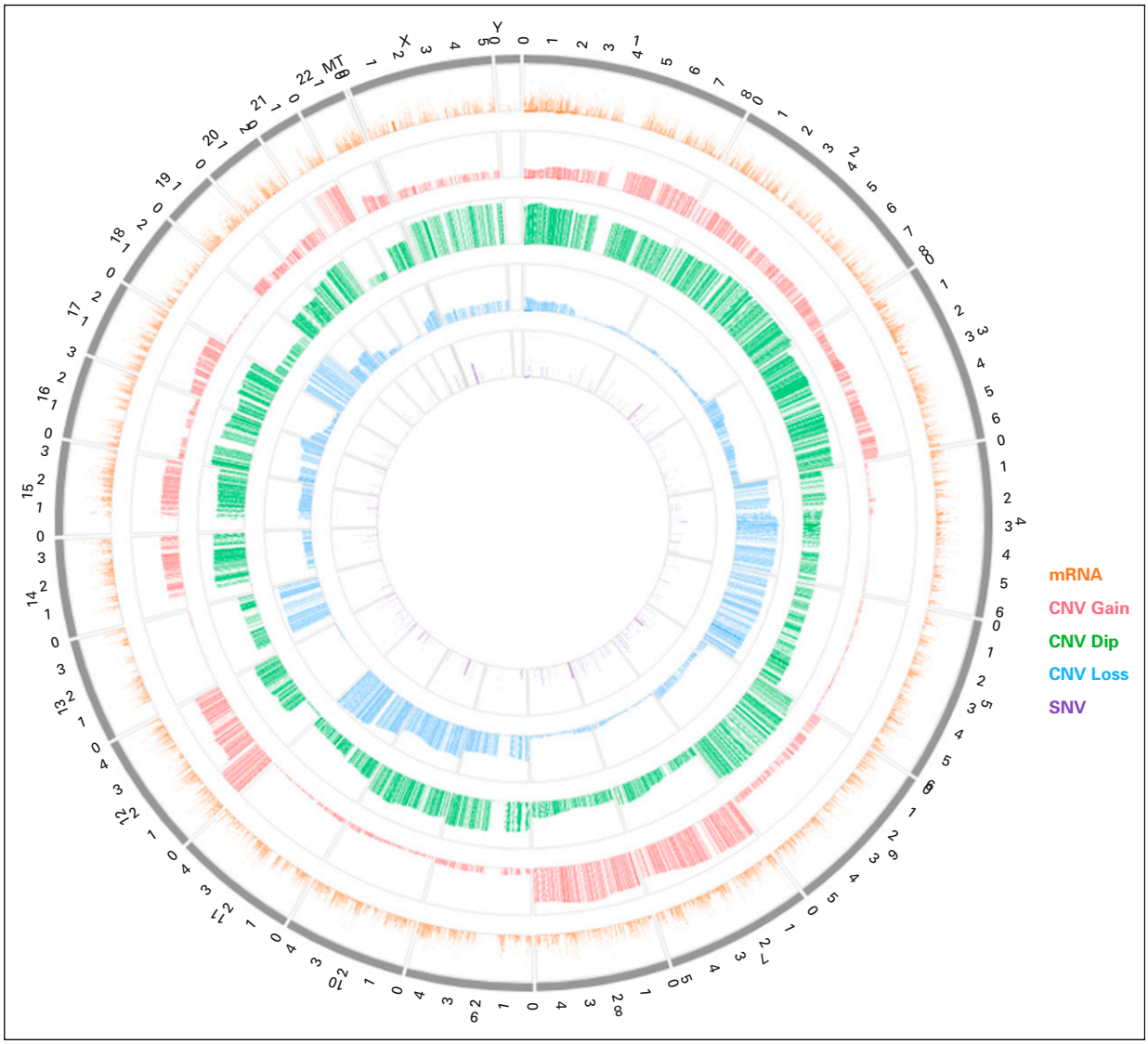


FIG 6. Representative visualization of high-level genomic data summary. The circos plot was generated for The Cancer Genomic Atlas Testicular Germ Cell Tumor data (N = 156). The circles from inside to outside stand for single-nucleotide variant (SNV) frequency (log transformed), copy-number variant (CNV) loss, CNV diploid (dip), CNV gain, mRNA expression, and chromosome location.

institutions. The utility and interoperability of the GCT data model can be greatly enhanced when connected to these generic EHR data models and exchange standards.

Currently, several data commons have been developed for cancers, such as the NCI GDC[22] and cBioPortal,[23] which have large advantages in their coverage of various types of disease, especially when a project requires cross-disease comparison. Compared with these general data commons, the GCT data commons is disease specific. It was built upon the GCT clinical model and contains important clinical variables. It aims to serve as a comprehensive GCT research resource by incorporating GCT data from MaGIC members, external data contributors, and the public domain. Another advantage of the GCT data commons is that all the integrated clinical data were curated according to the GCT data model, reducing time for data curation by the users. Furthermore, the visualization modules provided in the GCT data commons can meet users' needs for data exploration.

The current GCT data model covers the essential variables in each clinical episode, which match the extent of the currently available data sets. In the future, it will be expanded to capture more detailed information. For instance, radiotherapy is less frequently used for GCTs compared with surgery and chemotherapy, so the current GCT data model only records the start and end dates of radiotherapy. A future version is planned to record more detailed information, such as the type and dose of radiotherapy and treatment region. Moreover, we plan to expand the GCT specimen data model to include real-time sample tracking

and linking of genomic and imaging data to the clinical data. The current GCT data commons mainly focuses on clinical and genomic data. In the future, it will integrate digital images generated from GCT pathology slides and medical imaging modalities (eg, computed tomography and magnetic resonance imaging). It is noteworthy that current genomic analyses in pediatric GCTs have not yet been conclusive for real-world clinical use, partially because of the limited availability of genomic data sets. The GCT data model and data commons developed here may further facilitate the integration of genomic data from pediatric patients with GCTs and stimulate research in this area.

We developed the current GCT data model and data commons to overcome the sparsity and lack of interoperability of GCT data sets, which are mainly a function of the rarity of the disease. The current study is just the first step to develop data standards and data commons for GCTs. The data and input have been derived mainly from the pediatric GCT community, and we are in the process of integrating data sets from the postpubescent GCT community, which will be a critical next step to improve data sharing for the GCT community. The scale of this project and the availability of GCT data sets will grow as more researchers participate in this community effort by conforming and contributing their data according to the GCT data model. The current development focuses on pediatric GCTs, but the framework and experience of accomplishing this project, outlining clinical episodes, modeling data elements and variables, defining the controlled terminology, and constructing the data commons and visualization modules, can be relatively easily extended to postpubescent patients with GCTs. Beyond GCTs, the workflow and experience acquired in this project will serve as a learning opportunity for developing similar projects for other diseases, especially rare diseases.

AFFILIATIONS

[1]Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX
[2]Keck School of Medicine, University of Southern California, Los Angeles, CA
[3]Children's Oncology Group, Monrovia, CA
[4]Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX
[5]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX
[6]Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX
[7]Department of Pathology, University of Cambridge, Cambridge, United Kingdom
[8]Cancer and Blood Disease Institute, Children's Hospital Los Angeles, Los Angeles, CA
[9]Dana-Farber/Boston Children's Blood and Cancer Disorders Center, Boston, MA
[10]Hospital for Sick Children, University of Toronto, Toronto, ON, Canada
[11]Institute of Hematology and Pediatric Oncology, Lyon, France
[12]Department of Pediatric Oncology, Gustave Roussy, University of Paris-Saclay, Villejuif, France
[13]Center for Research Informatics, Division of Medicine and Biological Sciences, University of Chicago, Chicago, IL
[14]Department of Paediatrics, University College London Hospitals, London, United Kingdom
[15]Children's Cancer Hospital, Barretos Cancer Center, Barretos, Brazil
[16]Department of Paediatric Haematology and Oncology, Cambridge University Hospitals National Health Service Foundation Trust, Cambridge, United Kingdom

CORRESPONDING AUTHOR

Yang Xie, 5323 Harry Hines Blvd, ND3.101C, Dallas, TX 75390; e-mail: yang.xie@utsouthwestern.edu.

EQUAL CONTRIBUTION

B.C. and D.M.Y. contributed equally to this work as first authors, and A.L.F. and Y.X. contributed equally as senior authors.

SUPPORT

Supported by Grants No. 358099 from St Baldrick's Foundation, No. 5P30CA142543 and R01GM115473 from the National Institutes of Health, and No. RP180805 from the Cancer Prevention Research Institute of Texas.

AUTHOR CONTRIBUTIONS

Conception and design: Bo Ci, Donghan M. Yang, Qinbo Zhou, Stephen X. Skapek, Matthew J. Murray, James F. Amatruda, Samuel L. Volchenboum, James Nicholson, A. Lindsay Frazier, Yang Xie
Financial support: Yang Xie
Administrative support: Lindsay Klosterkemper, Yang Xie
Provision of study material or patients: James Nicholson, Yang Xie
Collection and assembly of data: Bo Ci, Mark Krailo, Caihong Xia, Bo Yao, Lin Xu, Matthew J. Murray, Lindsay Klosterkemper, Cecile Faure-Conter, Brice Fresneau, Sara Stoneham, Luiz Fernando Lopes, Yang Xie
Data analysis and interpretation: Bo Ci, Donghan M. Yang, Mark Krailo, Caihong Xia, Danni Luo, Guanghua Xiao, Lin Xu, Stephen X. Skapek, Matthew J. Murray, James F. Amatruda, Furqan Shaikh, Brice Fresneau, Samuel L. Volchenboum, Yang Xie
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Bo Ci
Employment: Genentech
Stock and Other Ownership Interests: Genentech
Travel, Accommodations, Expenses: Genentech

Mark Krailo
Consulting or Advisory Role: Merck Sharp & Dohme
Travel, Accommodations, Expenses: Merck Sharp & Dohme

Brice Fresneau
Consulting or Advisory Role: iQone Healthcare (Inst)

Samuel L. Volchenboum
Stock and Other Ownership Interests: Litmus Health
Honoraria: Sanford Health
Consulting or Advisory Role: CVS Accordant
Travel, Accommodations, Expenses: Sanford Health

A. Lindsay Frazier
Stock and Other Ownership Interests: Decibel Therapeutics
Consulting or Advisory Role: Decibel Therapeutics

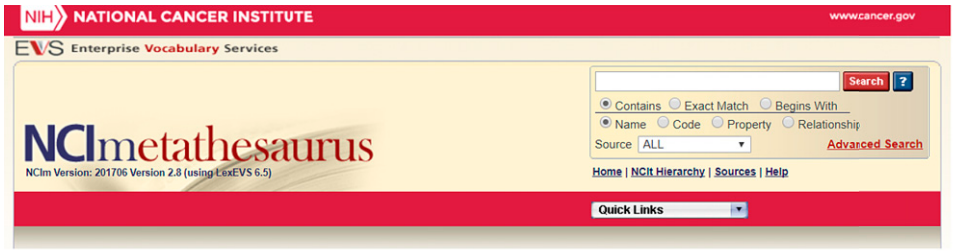
No other potential conflicts of interest were reported.

REFERENCES

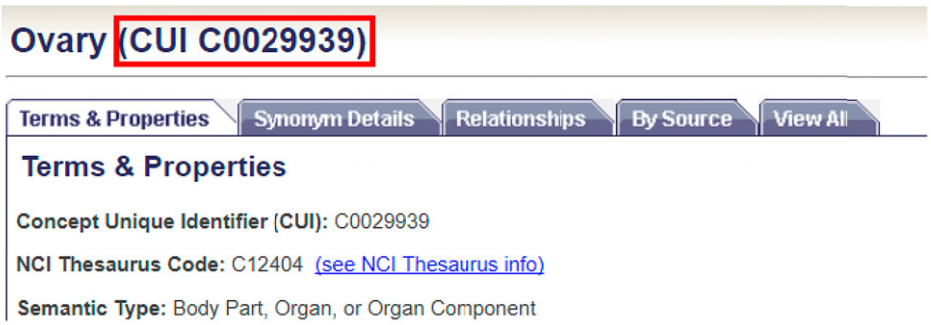
1. Oosterhuis JW, Looijenga LH: Testicular germ-cell tumours in a broader perspective. Nat Rev Cancer 5:210-222, 2005
2. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2019. CA Cancer J Clin 69:7-34, 2019
3. Göbel U, Schneider DT, Calaminus G, et al: Multimodal treatment of malignant sacrococcygeal germ cell tumors: A prospective analysis of 66 patients of the German cooperative protocols MAKEI 83/86 and 89. J Clin Oncol 19:1943-1950, 2001
4. Mann JR, Raafat F, Robinson K, et al: The United Kingdom Children's Cancer Study Group's second germ cell tumor study: Carboplatin, etoposide, and bleomycin are effective treatment for children with malignant extracranial germ cell tumors, with acceptable toxicity. J Clin Oncol 18:3809-3818, 2000
5. Cushing B, Giller R, Cullen JW, et al: Randomized comparison of combination chemotherapy with etoposide, bleomycin, and either high-dose or standard-dose cisplatin in children and adolescents with high-risk malignant germ cell tumors: A pediatric intergroup study—Pediatric Oncology Group 9049 and Children's Cancer Group 8882. J Clin Oncol 22:2691-2700, 2004
6. Lopes LF, Macedo CR, Pontes EM, et al: Cisplatin and etoposide in childhood germ cell tumor: Brazilian pediatric oncology society protocol GCT-91. J Clin Oncol 27:1297-1303, 2009
7. Duhil de Bénazé G, Pacquement H, Faure-Conter C, et al: Paediatric dysgerminoma: Results of three consecutive French germ cell tumours clinical studies (TGM-85/90/95) with late effects study. Eur J Cancer 91:30-37, 2018
8. Einhorn LH: Chemotherapeutic and surgical strategies for germ cell tumors. Chest Surg Clin N Am 12:695-706, 2002
9. Huddart RA, Norman A, Shahidi M, et al: Cardiovascular disease as a long-term complication of treatment for testicular cancer. J Clin Oncol 21:1513-1523, 2003
10. Travis LB, Ng AK, Allan JM, et al: Second malignant neoplasms and cardiovascular disease following radiotherapy. J Natl Cancer Inst 104:357-370, 2012
11. National Cancer Institute: Childhood extracranial germ cell tumor treatment. https://www.cancer.gov/types/extracranial-germ-cell/hp/germ-cell-treatment-pdq
12. Lopes LF, Macedo CR, Aguiar SS, et al: Lowered cisplatin dose and no bleomycin in the treatment of pediatric germ cell tumors: Results of the GCT-99 protocol from the Brazilian Germ Cell Pediatric Oncology Cooperative Group. J Clin Oncol 34:603-610, 2016
13. Fresneau B, Orbach D, Faure-Conter C, et al: Is alpha-fetopotein decline a prognostic factor of childhood non-seminomatous germ cell tumors? Results of the French TGM95 study. Eur J Cancer 95:11-19, 2018
14. Frazier AL, Hale JP, Rodriguez-Galindo C, et al: Revised risk classification for pediatric extracranial germ cell tumors based on 25 years of clinical trial data from the United Kingdom and United States. J Clin Oncol 33:195-201, 2015
15. National Institutes of Health: NIH commons overview, framework & pilots: Version 1. https://datascience.nih.gov/sites/default/files/CommonsOverview-FrameWorkandCurrentPilots281015_508.pdf
16. National Cancer Institute: NCIm. https://ncim-stage.nci.nih.gov/ncimbrowser/
17. Bagrodia A, Lee BH, Lee W, et al: Genetic determinants of cisplatin resistance in patients with advanced germ cell tumors. J Clin Oncol 34:4000-4007, 2016
18. Palmer RD, Murray MJ, Saini HK, et al: Malignant germ cell tumors display common microRNA profiles resulting in global changes in expression of messenger RNA targets. Cancer Res 70:2911-2923, 2010
19. National Center for Biomedical Computing: i2b2: Informatics for Integrating Biology and the Bedside. https://www.i2b2.org/
20. Klann JG, Phillips LC, Herrick C, et al: Web services for data warehouses: OMOP and PCORnet on i2b2. J Am Med Inform Assoc 25:1331-1338, 2018
21. Health Level Seven International: Fast Healthcare Interoperability Resources. http://hl7.org/fhir/
22. National Cancer Institute: Genomic Data Commons: Next Generation Cancer Knowledge Network. https://gdc.cancer.gov/
23. cBioPortal for Cancer Genomics: cBioPortal. https://www.cbioportal.org/

APPENDIX

A



B



C

D

Valid Value	NCIm term	Concept Unique Identifier (CUI)	Semantic Type	NCIt Definition
Ovary	Ovary	C0029939	Body Part, Organ, or Organ Component	One of the paired female reproductive glands containing the ova or germ cells; the ovary's stroma is a vascular connective tissue containing numbers of ovarian follicles enclosing the ova.
Testis	Testis	C0039597	Body Part, Organ, or Organ Component	Either of the paired male reproductive glands that produce the male germ cells and the male hormones.
Extragonadal	Extragonadal	C2986387	Body Location or Region	An area of the body other than the ovaries or testes.
Central Nervous System	Central Nervous System	C3714787	Body System	The part of the nervous system that consists of the brain, spinal cord, and meninges.
Mediastinum	Mediastinum	C0025066	Body Location or Region	A group of organs surrounded by loose connective tissue, separating the two pleural sacs, between the sternum anteriorly and the vertebral column posteriorly as well as from the thorac inferiorly. The mediastinum contains the heart and pericardium, the bases of the great vessels, the trachea and bronchi, esophagus, thymus, lymph nodes, thoracic duct, phrenic and vagus nerves, and other structures and tissues.
Retroperitoneum	Retroperitoneal Space	C0035359	Body Space or Junction	The back of the abdomen where the kidneys lie and the great blood vessels run. ic inlet superiorly to the diaphragm

FIG A1. National Cancer Institute Metathesaurus (NCIm) concept unique identifier (CUI) code mapping. (A) Screen copy of NCIm homepage. (B) Example CUI describing the term ovary. (C) List of sources (standards) that have the term ovary mapped to NCIm. (D) Mapping of six valid values in the germ cell tumor data model onto NCIm terms, CUIs, semantic types, and definitions according to the NCI Thesaurus (NCIt). AOD, Alcohol and Other Drug Thesaurus; CDISC, Clinical Data Interchange Standards Consortium; CSP, Computer Retrieval of Information on Scientific Projects (CRISP) Thesaurus; FMA, Foundational Model of Anatomy Ontology; ICDO, International Classification of Disease for Oncology; LNC, Logical Observation Identifiers Names and Codes; MDBCAC, Mitelman Database of Chromosome Aberrations in Cancer; MSH, Medical Subject Headings; MTH, Unified Medical Language System (UMLS) Metathesaurus; NCI-GLOSS, NCI Dictionary of Cancer Terms.

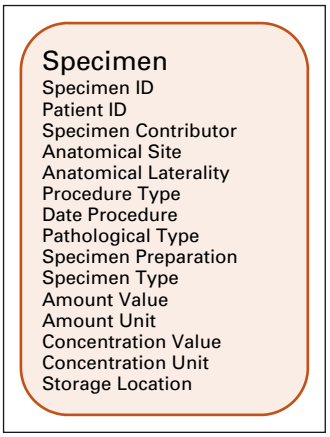


FIG A2. Representative variables in the germ cell tumor specimen data model. National Cancer Institute Metathesaurus (NCIm); NCIt, NCI Thesaurus.

