EJC Paediatric Oncology 4 (2024) 100186

Contents lists available at ScienceDirect

EJC Paediatric Oncology

journal homepage: www.journals.elsevier.com/ejc-paediatric-oncology

Common core variables for childhood cancer data integration

Daniela Di Carlo [a,b,*], Ruth Ladenstein [d], Norbert Graf [e,1], Johannes Hans Merks [f,g,2], Gustavo Hernández-Peñaloza [h,3], Pamela Kearns [c], Gianni Bisogno [a,b]

[a] Department of Women's and Children's Health, University of Padova, Italy
[b] Pediatric Hematology-Oncology Division, University Hospital of Padova, Italy
[c] Cancer Research UK Clinical Trials Unit, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham B15 2TT, UK
[d] St. Anna Children's Cancer Research Institute, Vienna, Austria; St Anna Children's Hospital, Vienna, Austria
[e] Saarland University, Dep. Paediatric Oncology and Haematology, Homburg 66421, Germany
[f] Princess Máxima Center for Pediatric Oncology, Utrecht, the Netherlands
[g] Division of Imaging and Oncology, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands
[h] Universidad Politécnica de Madrid, ETSISI, Departamento de Sistemas informáticos, Madrid, Spain

ARTICLE INFO

Keywords:
Harmonisation
Big-data
Paediatric oncology
AI
Health research

ABSTRACT

Introduction: Data-driven research has improved paediatric cancer outcomes for children. However, challenges in sharing data between institutions prevent the use of artificial intelligence (AI) to address substantial unmet needs in children diagnosed with cancer. Harmonising collected data can enable the application of AI for a greater understanding of paediatric cancers. The main goal of the paper was to analyse the currently used childhood cancer databases to identify a core of variables able to capture the most relevant data on the diagnosis and treatment of children and adolescents with cancer.

Methods: We arbitrarily identified different types of existing databases dedicated to collecting data of patients with solid tumours, Umbrella, FAR-RMS; PARTNER; ERN PAEDCAN Registry; INSTRUCT and INRG; the common data elements for Rare Disease by Joint Research Centre. The different elements of the CRFs were analysed and ranked "essential" and "good to have". Domains that included a group of variables structurally connected were identified. Each variable was defined by name, data type, description, and permissible values.

Results: We identified six structural domains: Patient registration, Personal information, Disease History, Diagnosis, Treatment, and Follow-up and Events. For each of them, "essential" and "good to have" variables were defined.

Discussion: Data harmonisation is essential for enhancing integration and comparability in research. By standardizing data formats and variables, researchers can facilitate data sharing, collaboration, and analysis across multiple studies and datasets. Embracing data harmonization practices will advance application of AI, scientific knowledge, improve research reproducibility, and contribute to evidence-based decision-making in various fields.

1. Introduction

The advancement in child and adolescent cancer survival is a significant achievement in paediatrics, largely due to collaborative efforts across national and international networks dedicated to clinical and biological research [1]. European cooperative groups, unified under the International Society for Paediatric Oncology (SIOPE),[4] have conducted trials since the 1970s, resulting in extensive data collection. However, the data structure and technology variability pose a challenge for modern big data analysis tools [2].

To address this, the UNICA4EU (Towards a UNIque approach for artificial intelligence data-driven solutions to fight Childhood cAncer

* Correspondence to: via Giustiniani 1, Padova 35128, Italy
E-mail address: daniela.dicarlo@aopd.veneto.it (D. Di Carlo).
[1] ORCID: 0000–0002-2248–323X
[2] ORCID: 0000–0001-7659–1028
[3] ORCID: 0000–0003-2177–6185
[4] Available online: https://siope.eu/european-research-and-standards/clinical-trials-in-paediatric-oncology/

https://doi.org/10.1016/j.ejcped.2024.100186
Received 30 April 2024; Received in revised form 9 August 2024; Accepted 21 August 2024
Available online 31 August 2024

FOR Europe)[5] project aims to establish standardised core variables for childhood cancer research, facilitating data comparison and collaboration. These variables encompass demographic, diagnostic, treatment, and follow-up information. UNICA4EU, coordinated by SIOPE, involves 15 partners across Europe and aligns with the European Beating Cancer Plan Flagship's initiative 'Helping Children with Cancer Initiative'[6]; to improve childhood cancer detection, treatment, and care focused on ensuring that children have access to rapid and optimal detection, diagnosis, treatment, and care and building cancer centre networks that extend and complement the European Reference Network for Paediatric Cancer (ERN PaedCan[7]). (see Supplemental material Table 1).

The project collaborates with stakeholders like Childhood Cancer International (CCI)-Europe,[8] PanCare,[9] and ERN PaedCan to bridge technical research and policy advocacy for AI adoption in paediatric oncology.

The paper's main objective is to propose a coding system for future database implementation, aiming to optimize AI tool utilization despite data acquisition challenges.

2. Methods

We arbitrarily identified different types of existing databases dedicated to collecting data of patients with solid tumours. As a proof of concept and to explore different data sources, we decided to select for this analysis the databases of a. two ongoing clinical trials, Umbrella (EudraCT Number 2016–004180–39) and FAR-RMS (Frontline and Relapse RhabdoMyoSarcoma Study, ClinicalTrials.gov Identifier: NCT04625907, EudraCT Number 2018–000515–24); b. a prospective observational study aiming at registering children with very rare tumours, PARTNER (Paediatric Rare Tumours Network) [3]; c. a data dictionary produced within the ERN PAEDACAN initiative, ERN PAEDCAN Registry[10]; d. two global databases where an effort of harmonization has already been undertaken at an international and transatlantic level, INSTRUCT (INternational Soft Tissue SaRcoma ConsorTium[11]) and INRG (International Neuroblastoma Risk Group[12]), e. the common data elements for Rare Disease Registration released by the European Union Rare Disease Platform, Joint Research Centre (JRC).[13]

Table 1
Domain's characteristics.

	Domain name	Description
1	**Patient registration**	It contains information on patient identification including a code and the centre of registration.
2	**Personal information**	It collects patient information like sex, age, and pre-existing diseases.
3	**Disease History**	It contains the relevant information from the onset of the symptoms to the disease diagnosis.
4	**Diagnosis**	It collects information about the diagnosis, including the type of diagnosis, when and how (biopsy/radiology) has been made and the extension of the disease.
5	**Treatment**	It summarizes the type of treatment, including chemotherapy, radiotherapy, and surgery.
6	**Follow-up and Events**	It describes patient status at the last follow-up and relevant events after the end of treatment including relapse or sequelae

The SIOP Renal Tumour Study Group introduced the Umbrella Protocol for childhood renal tumours, encompassing treatment guidelines for Wilms tumours and other renal neoplasms. With a comprehensive database comprising 1298 items, patient registration began in 2019 and is still ongoing [4]. The protocol evaluates the prognostic significance of genomic alterations within the tumour and the remaining blastemal component post-preoperative chemotherapy.

The FaR-RMS study registers children and adults with newly diagnosed and relapsed rhabdomyosarcoma (RMS). It is a multi-arm, multi-stage trial involving several randomised chemotherapy and radiotherapy questions in both newly diagnosed and relapsed RMS. The study was opened in September 2020, and it is ongoing.

The peculiarity of UMBRELLA and FAR RMS for the purpose of our analysis are that they are databases built to collect basic data but especially specific data to answer trial questions.

The PARTNER study seeks to establish a European prospective registry for children and adolescents with very rare tumours (VRT). The database, created by harmonising variables from various EU national registries, began recruitment in 2023 in specific countries. PARTNER's variables are accessible through the European Rare Disease Registry Infrastructure's Central Metadata Repository.[14]

The ERN PaedCan Registry, part of the European Reference Networks (ERNs), aims to foster collaboration among national health systems to benefit patients. Specifically, ERN PaedCan strives to enhance childhood cancer survival rates by offering equitable, high-quality cross-border care. It provides a standardised core dataset and 15 disease-specific datasets, aiding countries in shaping their national databases. This registry results from efforts to harmonise variables from diverse protocols and datasets.

The INRG task force started its activity in 2004. A database was established to collect data of patients with neuroblastoma registered in studies conducted by different cooperative groups: Children's Oncology Group (COG),[15] Gesellschaft für Pädiatrische Onkologie und Hämatologie (GPOH),[16] Neuroblastoma Committee of Japan Children's Cancer Group (JCCG),[17] Japanese Infantile Neuroblastoma Co-operative Study Group (JINCS), and Society of Paediatric Oncology Europe Neuroblastoma Group (SIOPEN). The initial set included 8800 patients, but the amount of data continues to expand, as the cooperative groups have agreed to update existing patient data and add new patient data once clinical trials are completed and the objectives of the trial are published.

INSTRuCT, founded by Children's Oncology Group (COG), Cooperative Weichteilsarkom Studiengruppe (CWS), and European paediatric Soft tissue sarcoma Study Group (EpSSG), aims to create a unified international risk stratification system for RMS, replacing disparate systems in Europe and North America. It also incorporates data on other soft tissue sarcomas. With information from over 7000 patients enrolled in previous STS trials, secondary analyses are planned, leveraging the dataset's size for rare tumour research [5]

[5] Available online: https://unica4.eu/
[6] Available online: https://health.ec.europa.eu/non-communicable-diseases/cancer/flagship-initiatives_en
[7] Available online: https://paedcan.ern-net.eu/
[8] Childhood Cancer International – Europe. CCI-E. Available online: https://ccieurope.eu/
[9] PanCare network. Available online: https://www.pancare.eu/
[10] ERN PAEDCAN Registry. Available online: https://www.escp-registry.org/myhealth/
[11] International Soft Tissue Sarcoma Consortium. Available online: https://datacatalog.ccdi.cancer.gov/dataset/PCDC-INStruCt
[12] International Neuroblastoma Risk Group. Available online: https://inrgdb.org/
[13] Available online: https://eu-rd-platform.jrc.ec.europa.eu/_en
[14] Available online: ERDRI, https://eu-rd-platform.jrc.ec.europa.eu/erdri_en
[15] Children's Oncology Group COG: https://www.childrensoncologygroup.org/
[16] https://www.gpoh.de/
[17] Available online: http://jccg.jp/en/

Both INRG and INSTRuCT represent a multilateral collaboration, and databases have been built with the support of Chicago Data Commons.[18] This required a preliminary analysis of the data (including variables and coding) collected in the databases managed by the contributing Cooperative Groups. Moreover, the important harmonisation effort from both initiatives forms the basis for merged analysis, which is of special interest to our analysis.

The "Set of Common Data Elements for Rare Diseases Registration"[19] was developed by a JRC-coordinated Working Group, integrating expertise from EU projects like EUCERD (European Union Committee of Experts on Rare Diseases),[20] EPIRARE (European Platform for Rare Disease Registries),[21] and RD-Connect. It standardises 16 essential data elements for rare disease registries, including paediatric cancer, to enhance interoperability and facilitate research.

These protocols and registries were analysed using case report forms (CRFs). The different elements of the CRFs were transferred into an Excel sheet and aligned so that similar elements were placed next to each other. Understanding the number of parameters to feed the dataset for potential use of AI is a crucial step to design effective AI models. Increasing the number of parameters can enhance the model's ability to comprehend complex data patterns, potentially improving accuracy. However, it is essential to find the right balance. An excessive number of parameters can cause the model to memorize specific examples from its training data rather than learning the general patterns. This overfitting can lead to poor performance on new, unseen data. Thus, limiting the number of parameters is crucial in model development. For this reason, a ranking system based on two tiers was deemed necessary to define which variables are mandatory to describe the patient, the tumor, the treatment, and its results ("essential"). Some variables which complete the information but are not strictly necessary to understand the story of the patient and of the tumour, and that are likely to be collected have been defined as "good to have". Variables were classified as "essential" or "good to have" through consensus among investigators (GB, DDC), then presented to the UNICA4EU partners and discussed during the project meetings to achieve a consensus. Additionally, domains grouping structurally connected variables were identified. Each variable was defined with standardised names, data types, descriptions, and values using NCI terminologies for data replicability and interoperability.[22]

3. Results

We identified six structural domains: 1) Patient registration, 2) Personal information, 3) Disease History, 4) Diagnosis, 5) Treatment, and 6) Follow-up and Events. The Patients Registration domains are depicted in Table 1.

3.1. Domain 1. Patient registration

3.1.1. Essential variables

After thoroughly comparing the different datasets, a set of essential variables has been identified.

- *Subject ID:* A vital variable present across all databases, albeit with varying names, is essential for patient identification within each trial. Pseudonymization, as recommended by the JRC, ensures patient anonymity, particularly in paediatric oncology. Variables like Centre and Country of Registration were included in three databases, crucial for understanding epidemiology, treatment, and outcomes across different countries. However, INRGand INSTRuCT, gathering data from existing databases, did not incorporate these variables.
- *Study ID:* This information was present only in INRG and INSTRuCT, but it seems important to have it for potential future data exchange.
- *Informed consent:* The treating physician is responsible for obtaining informed consent to collect data, but it is important to confirm that this has been done and include this information in the database.

3.1.2. Good to have variables

Other variables that were present at least in some of the databases analysed but were considered not essential are:

- *Country of registration;*
- *Centre of registration;*
- *Study arm;*
- *To be contacted;*
- *To reuse data.*

The following variables were not included because of privacy issues:

- *Family name;*
- *Given name.*

3.2. Domain 2. Personal information

3.2.1. Essential variables

- *Year of birth:* Collecting patients' dates of birth is optimal for research purposes, but it conflicts with privacy regulations. Two datasets (PARTNER, Umbrella) collected this information, aligning with JRC recommendations, while three others substituted it with age at diagnosis. Date of birth enhances patient identifiability, especially in rare diseases, but age at diagnosis offers privacy compliance. Nonetheless, collecting birth years allows for contextualizing patients over time, facilitating analysis of outcome and prognosis changes across decades due to therapy advancements.
- *Sex:* The variable "sex" reflects an individual's biology, while "gender" denotes gender identity. This offers dual advantages for research and considering risk factors, especially for ongoing medical therapies in adolescents and young adults. While gender was not included in datasets except for the Umbrella protocol, where it refers to sex, we've added it as a non-essential variable.
- *Pre-existing medical conditions:* This variable should describe information regarding pre-existing medical conditions, including malformations or previous tumours.

3.2.2. Good to have variables

In addition to essential variables, some databases include non-essential ones like race, country of birth, family history, number of siblings, assisted reproductive fertilisation, birth weight, gestational age at birth, performance status (Karnofsky), and congenital abnormalities. However, including race raises complex issues due to its historical and scientific implications. While it may provide epidemiologically relevant data, its use in genetic research is controversial. To address this, scientific communities should promote terms like "ancestry" or "population" and require authors to clarify their definitions and usage. Despite privacy and ethical concerns, collecting race or ancestry data could help identify disparities in treatment access and tailor support for marginalised groups. Moreover, race-related biological differences could impact disease biology and drug metabolism, emphasizing the need for inclusive research practices.

[18] Chicago Data Commons. Available online: https://ctds.uchicago.edu/data-commons

[19] Available online: https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en

[20] EUCERD Joint Action. Available online: https://www.eunethta.eu/eucerd/

[21] EPIRARE. Available online: https://www.raredis.org/archives/1200?lang=en

[22] NCI Terminology: https://www.cancer.gov/publications/dictionaries

3.3. Domain 3. Disease history

3.3.1. Essential variables

This domain summarizes the medical history from the onset of symptoms to the diagnosis, it is mentioned in PARTNER, Umbrella and JRC Registry, while Far-RMS, INSTRUCT_RMS and INRG do not include it. We did not identify essential variables in this domain, but "good to have" variables may be interesting for epidemiological analysis.

3.3.2. Good to have variables

Other variables that could be interesting for a better understanding of the disease history may be challenging to obtain or standardise. Therefore, the following entries have been considered non-essential but "good to have": age at onset of symptoms, symptoms and signs, route to diagnosis, duration of symptoms, and undiagnosed case.

Special considerations should be made for the following entries:

- *Age at first contact with specialised centre:* Similar to the duration of symptoms, it may suggest the interval between the appearance of symptoms of the disease and the subsequent referral to a specialised paediatric oncology centre, emphasising potential challenges in accessing specialised care, typical of extensive regions or areas with an ineffective hospital network. This aspect is pertinent to studies concerning equality of access to care and its impact on the final outcome.
- *Patient type:* This information is necessary if the database registers patients at new diagnoses or at the moment of relapse. It is collected in INRG and Umbrella protocol datasets because in this system, every event requires the registration of a new form. Our recommendations replace it with the details contained in the follow-up domain, which describes any events during the course of the disease, including relapses.

3.4. Domain 4. Diagnosis

3.4.1. Essential variables

- *Diagnosis: T*he International Classification of Childhood Cancer, Third Edition (ICCC-3)[6], is a widely accepted system for diagnosing cancer in paediatric oncology. It provides a comprehensive framework for classifying and coding childhood cancers based on morphology, topography, and behaviour. ICCC-3 is recommended by international organisations and is specifically designed for paediatric cancer classification. It helps standardise cancer reporting, enabling better comparability of data across different regions and institutions. Using ICCC-3 for coding cancer diagnoses in paediatric oncology is generally considered a suitable and standardised approach. We observed that among the various protocols, only the PARTNER Registry recommends using ICCC-3 in such cases. Conversely, other protocols tailored to specific disease types already advise identifying the subtype of histological diagnosis at this stage.
- *Molecular abnormalities:* They are collected in PARTNER, but also in Umbrella (predisposing mutations) and Far-RMS (fusion-status), in different sections. This is often a specific tumour information and coding, and the list of abnormalities needs to be adapted accordingly.
- *Age at diagnosis:* Alongside the year of birth, the age at onset plays a crucial role in contextualising the patient over time. When combined with the age at remission, relapse, death, or experiencing toxicity (collected in the subsequent domains), it becomes a valuable tool for outcome analysis, contributing to the definition of disease-free survival (DFS), event-free survival (EFS), and overall survival (OS). To report the total number of months seems more practical than asking for years and months.
- *Site:* In paediatric oncology, the International Classification of Diseases for Oncology, Third Edition (ICD-O-3) is commonly used for coding cancer sites.[23] ICD-O-3 provides a comprehensive system for classifying and coding neoplasms, including morphology and topography codes. It is widely accepted and used globally for cancer registration. ICD-O-3 allows for detailed coding of tumour characteristics, aiding in accurate and standardised cancer data reporting. The morphology code specifies the type of cancer cells, while the topography code indicates the organ or tissue where the cancer originates. For paediatric oncology, where accurate and detailed coding is crucial, ICD-O-3 is generally recommended. ICD-O-3 is used in PARTNER Registry, while, similarly to "Diagnosis", in the other protocols there is a list of sites typically involved for the specific diseases, not exhaustive when considering all the tumour entities.
- *Tumour local Invasiveness, Nodal involvement, Metastases:* Despite its simplicity compared to modern diagnostic tools, the TNM classification system serves as a potential framework in paediatric oncology for summarizing disease extent comprehensively. While it may seem rudimentary, TNM offers a valuable means to encapsulate disease spread across different diagnostic categories. Notably, specifying lymph node involvement based on imaging is crucial, as not all cases undergo surgical confirmation. However, histologically confirmed lymph node invasion is considered "good to have," highlighting the need for a nuanced approach to TNM's utility and limitations.

TNM classification is commonly used in paediatric oncology settings, including the PARTNER registry, Far-RMS protocol, and INSTRuCT. However, supplementing TNM with a more tumour-specific staging system, as recommended by the Umbrella protocol and INRG, may be necessary to better tailor staging to specific conditions.

- *Tumour size*: this may be part of the TNM staging system and categorized as a or b.
- *Sites of metastasis*: a list of the most common sites of metastasis is provided in the dataset.

3.4.2. Good to have variables

Central review; biological sample available for research; paraffin tumour tissues; frozen tumour; patient blood; molecular abnormalities; Disease status; Nodal pathology.

3.5. Domain 5. Treatment

3.5.1. Essential variables

- *Inclusion in a national or international treatment trial/protocol:* Due to its overarching nature, this information is only collected in PARTNER. The other considered datasets already represent a national/international protocol. It is an important variable to have in performing analysis across multiple databases.
- *Specify which trial or protocol:* This information is only collected in PARTNER but should be added for the same reasons explained above.
- *Treatment line:* This information, similarly to the previous one, is only collected in PARTNER.
- *Standard chemotherapy (start and end date):* This information is easy to collect and uniformly collected in all the considered datasets. The minimum required information includes whether standard chemotherapy has been delivered and from when to when. In this case, the

[23] WHO. International Classification of Diseases for Oncology. Available online: https://iris.who.int/bitstream/handle/10665/96612/9789241548496_eng.pdf

dates collected give a clue about the length of the therapy without compromising the individual's privacy.

- *Radiation therapy administered (start and end date), Type and dose of radiotherapy:* This easily obtainable information is consistently gathered across all the datasets under consideration. The essential details encompass whether standard radiotherapy was administered, the duration, and the type of energy employed.
- *The surgical procedure, Age and Date of surgery on the primary tumour, and Surgical outcome:* This information is also uniformly collected in all datasets. It is essential to determine whether surgery has been performed, when it occurred, and whether it was radical or not according to the R0, R1 and R2 system.
- *Best response, Age at end of treatment, Disease status at the end of therapy:* Each dataset consistently records a comprehensive evaluation of the disease post-therapy and the patient's current status.
- *Immunotherapy (start and end date); Targeted therapy (start and end date):* Immunotherapy and targeted therapy, while not presently integral to standard protocols, are increasingly utilised, particularly in cases of rare tumours or relapse. Despite the absence of information about these therapies in the considered datasets, we believe it is appropriate to gather such data as it enhances the understanding and completeness of the patient's treatment journey. The essential information required includes whether these therapies were administered. Given the diversity of drugs, knowing the specific compound used and any associated side effects would be beneficial. These additional details are categorised under the "good to have" section.

3.5.2. Good to have variables

Drugs (chemotherapy); Major adverse effects (due to chemotherapy, grade IV); Major adverse effect (due to radiotherapy, grade IV); Target area of radiotherapy; Drugs (immunotherapy); Major adverse effects to immunotherapy (grade IV); Drugs (Targeted therapy); Major adverse effects to targeted therapy (grade IV).

3.6. Domain 6. Follow up and events

3.6.1. Essential variables

- *Age at follow up, Age at event, Age at death:* Alongside the year of birth and the age at onset, they play a crucial role in contextualising the patient over time. When combined with the age at remission, relapse, death, or experiencing toxicity (collected in the subsequent domains), it becomes a valuable tool for outcome analysis, contributing to the definition of disease-free survival (DFS), event-free survival (EFS), and overall survival (OS).
- *Patient's status at last follow-up, An event has occurred, Type of event, Cause of death, Presence of sequelae, In case of a second tumour please specify the type, Treatment of the event, Type of treatment of the event:* This final section garnered the highest level of agreement among the datasets, primarily because this information is consistently required. Its uniform necessity, ease of summarization, and standardization make it valuable, offering details that contribute significantly to comprehending the disease's history.

3.6.2. Good to have variables

- More details about the treatment of the event.

4. Discussion

Data harmonisation refers to combining data from different sources and providing users with a comparable view of data from different studies. This process is becoming increasingly significant in demography and sociology research since the need for data harmonisation is rapidly growing as the volume and the need to share existing data explode [7] [8] [9].

In adult oncology, extensive data utilisation is expected due to large patient populations and established registries like the US Genomic Data Commons (GDC). However, with fragmented initiatives across European countries, paediatric oncology faces more challenges. Selected studies like Umbrella, FaR-RMS, PARTNER, INRG, and INSTRUCT showcase efforts to collect crucial clinical and biological data for specific inquiries. While not exhaustive of paediatric oncology databases, these studies demonstrate diverse variability and international collaboration.

Harmonisation in data collection involves standardising terminology, categories, formats, and methods, leading to more insightful and valuable statistical outputs. This enhances users' understanding and effectively meets their needs. Additionally, it reduces costs by eliminating duplication efforts [10].

Establishing a core set of variables in paediatric oncology fosters consistent data collection, promoting collaboration across trials and registries. Harmonisation facilitates data pooling, advancing our understanding of childhood cancers and improving treatment outcomes. This work marks the initial harmonisation phase among EU paediatric oncology groups, aiming to aggregate high-quality data for AI utilisation. While harmonised categories streamline big data analysis, the process faces barriers and challenges.

Brain tumour, leukaemia, and lymphoma protocols were not sources for the dataset, posing challenges in summarising relevant information, especially for leukaemia. Liquid tumours like leukaemia require distinct variables to capture their unique characteristics and treatment dynamics. A specialised core of variables is essential to ensure accurate representation in research focused on these conditions.

Data dictionaries exhibited variations in variable names, types, and coding systems, posing challenges in compiling an optimal core. Efforts prioritised standardisation and avoided subjective descriptions, aiming for coherence and reliability despite potential rigidity. Balancing comprehensive tumour comparisons with disease-specific information is crucial for nuanced understanding without overwhelming data collection. Striking this balance optimises AI capabilities while avoiding data excess or insufficiency. Sustainability poses a challenge, particularly with evolving classifications like ICDO3 and changing items over time. Selecting enduringly relevant elements helps at our best mitigate data dictionary ageing, ensuring ongoing utility and accuracy.

The exclusion of molecular and imaging data limits the dataset's scope, overlooking essential information for paediatric oncology research. Incorporating radiomics and radiogenomics could enhance understanding and should be considered in future iterations[11].

In conclusion, data harmonisation is vital for research integration and comparability. Standardising variables facilitates collaboration and analysis across studies. However, effective harmonisation requires careful planning and consideration of ethical and legal aspects. Embracing harmonisation practices advances AI applications, enhances research reproducibility, and supports evidence-based decision-making.

Funding

UNICA4EU has received funding from the European Union's Call for Pilot Projects and Preparatory Actions (PPPA) under Grant Agreement No 101052609

CRediT authorship contribution statement

Ruth Ladenstein: Supervision, Methodology, Conceptualization. **Norbert Graf:** Writing – review & editing, Validation, Supervision, Methodology, Data curation, Conceptualization. **Johannes Hans Merks:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Gustavo Hernández-Peñaloza:** Writing – review & editing, Validation. **Pamela Kearns:** Supervision, Project administration, Methodology, Conceptualization. **Gianni Bisogno:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology,

Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We deeply thank all the partners involved in UNICA4EU project, especially who participated in the discussions to improve our work. We thank: Schneider Carina, Bocolli Albina (CHILDHOOD CANCER INTERNATIONAL EUROPE (CCI-E)); Ladenstein Ruth, Planinic Adriana, Dobai Zoltan, Mancini Serena (ST. ANNA KINDERKREBSFORSCHUNG GMBH (CCRI)); Cattaneo Carlotta, Bicchieri Marilena, Maffia Fiore, Petroni Danilo, Pipolo Simona, Meoni Massimiliano (HUMANITAS MIRASOLE (ICH)); Gangas Pilar, Alvarez Arturo, Ferrer Georgina, Perrin Martin, Lyne Fiona (INTERNATIONAL FOUNDATION FOR INTEGRATED CARE (IFIC)); Lecinse Carole, Gilles Vassal (INNOVATIVE THERAPIES FOR CHILDREN WITH CANCER ASSOCIATION (ITCC)); Tozzi Alberto, Dell'Anna Vito Andrea (OSPEDALE PEDIATRICO BAMBINO GESÙ (OPBG)); Scheinemann Katrin, van der Pal Heleen, te Dorsthorst Jeroen (PAN-EUROPEAN NETWORK FOR CARE OF SURVIVORS AFTER CHILDHOOD AND ADOLESCENT CANCER (PanCare)); Merks Hans, Heuvel-Eibrink, M.M. van den (Marry), Krijger-2, R.R. de (Ronald), Schoot Reineke, Borja Jimenez Karina, Kemmeren Patrick (PRINSES MAXIMA CENTRUM VOOR KINDERONCOLOGIE BV (PMC)); Meyerheim Marcel, Graf Norbert, Shuping Wen, Theis Claudia (SAARLAND UNIVERSITY (USAAR)); Kearns Pamela, Nikki Coleman, Rizzari Carmelo, Essiaf Samira, Torou Elena, De Mofftards Vinciane (SIOPE); Antoniotti Marco, Donato Alessia, Crespi Federica, Masiero Laura (UNIVERSITA' DEGLI STUDI DI MILANO-BICOCCA (UNIMIB)); Battini Sara, Virgone Calogero (UNIVERSITA DEGLI STUDI DI PADOVA (UNIPD)); Álvarez Federico, Hernandez Gustavo (UNIVERSIDAD POLIÉTNICA DE MADRID (UPM)); Nefria Begonya, Royo Victor (FUNDACIO PRIVADA PER A LA RECERCA ILA DOCENCIA SANT JOAN DE DEU (FSJD)). Moreover, we thank Samuel L. Volchenboum and his staff, Meriel Jenney and Far RMS managing staff.

Approval

Due to the nature of the study not involving patient data, a review board approval was not requested

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejcped.2024.100186.

References

[1] M.J. Ehrhardt, K.R. Krull, N. Bhakta, Q. Liu, Y. Yasui, L.L. Robison, et al., Improving quality and quantity of life for childhood cancer survivors globally in the twenty-first century, Nat. Rev. Clin. Oncol. 20 (10) (2023) 678–696.
[2] N. Bhakta, L.M. Force, C. Allemani, R. Atun, F. Bray, M.P. Coleman, et al., Childhood cancer burden: a review of global estimates, Lancet Oncol. 20 (1) (2019) e42–e53.
[3] D. Orbach, A. Ferrari, D.T. Schneider, Y. Reguerre, J. Godzinski, E. Bien, et al., The European paediatric rare tumours network - European registry (PARTNER) project for very rare tumors in children, Pediatr. Blood Cancer 68 (Suppl 4) (2021) e29072.
[4] M.M. van den Heuvel-Eibrink, J.A. Hol, K. Pritchard-Jones, H. van Tinteren, R. Furtwängler, A.C. Verschuur, et al., Position paper: rationale for the treatment of Wilms tumour in the UMBRELLA SIOP-RTSG 2016 protocol, Nat. Rev. Urol. 14 (12) (2017) 743–752.
[5] D.S. Hawkins, G. Bisogno, E. Koscielniak, Introducing INSTRuCT: an international effort to promote cooperation and data sharing, Pediatr. Blood Cancer 70 (3) (2023) e28701.
[6] E. Steliarova-Foucher, C. Stiller, B. Lacour, P. Kaatsch, International classification of childhood cancer, third edition, Cancer 103 (7) (2005) 1457–1467.
[7] G. Weiler, U. Schwarz, J. Rauch, K. Rohm, T. Lehr, S. Theobald, et al., XplOit: an ontology-based data integration platform supporting the development of predictive models for personalized medicine, Stud. Health Technol. Inf. 247 (2018) 21–25.
[8] M. Cases, L.I. Furlong, J. Albanell, R.B. Altman, R. Bellazzi, S. Boyer, et al., Improving data and knowledge management to better integrate health care and research, J. Intern. Med. 274 (4) (2013) 321–328.
[9] P. Knaup, S. Garde, A. Merzweiler, N. Graf, F. Schilling, R. Weber, et al., Towards shared patient records: an architecture for using routine data for nationwide research, Int. J. Med. Inf. 75 (3–4) (2006) 191–200.
[10] Auffray, C. Balling, R. Barroso, I. Bencze, L. Benson M, J. Bergeron, et al., Making sense of big data in health research: towards an EU action plan, Genome Med. 8 (1) (2016) 71.
[11] M. Tsiknakis, V.J. Promponas, N. Graf, M.D. Wang, S.T.C. Wong, N. Bourbakis, et al., Guest editorial: computational solutions to large-scale data management and analysis in translational and personalized medicine, IEEE J. Biomed. Health Inf. 18 (3) (2014) 720–721.