Knowledge Representation of a Multicenter Adolescent and Young Adult Cancer Infrastructure: Development of the STRONG AYA Knowledge Graph

Joshi Hogenboom, MSc[1]; Varsha Gouthamchand, PhD[1]; Charlotte Cairns, MSc[2,3]; Silvie H.M. Janssen, PhD[4,5,6]; Kirsty Way, MSc[2]; Andre L.A.J. Dekker, PhD[1]; Winette T.A. van der Graaf, MD, PhD[4,5]; Anne-Sophie Darlington, PhD[2]; Olga Husson, PhD[4,7]; Leonard Y.L. Wee, PhD[1]; Johan van Soest, PhD[1,8]; and Aiara Lobo Gomes, PhD[1,9]

DOI https://doi.org/10.1200/CCI-25-00177

ABSTRACT

PURPOSE Rare diseases are difficult to fully capture, and regularly call for large, geographically dispersed initiatives. Such initiatives are often met with data harmonization challenges. These challenges render data incompatible and impede successful realization. The STRONG AYA project is such an initiative, specifically focusing on adolescent and young adult (AYAs) with cancer. STRONG AYA is setting up a federated data infrastructure containing data of varying format. Here, we elaborate on how we used health care–agnostic semantic web technologies to overcome such challenges.

METHODS We structured the STRONG AYA case-mix and core outcome measures concepts and their properties as knowledge graphs. Having identified the corresponding standard terminologies, we developed a semantic map on the basis of the knowledge graphs and the here introduced annotation helper plugin for *Flyover*. *Flyover* is a tool that converts structured data into resource description framework (RDF) triples and enables semantic interoperability. As a demonstration, we mapped data that are to be included in the STRONG AYA infrastructure.

RESULTS The knowledge graphs provided a comprehensive overview of the large number of STRONG AYA concepts. The semantic terminology mapping and annotation helper allowed us to query data with incomprehensible terminologies, without changing them. Both the knowledge graphs and semantic map were made available on a Hugo webpage for increased transparency and understanding.

CONCLUSION The use of semantic web technologies, such as RDF and knowledge graphs, is a viable solution to overcome challenges regarding data interoperability and reusability for a federated AYA cancer data infrastructure without being bound to rigid standardized schemas. The linkage of semantically meaningful concepts to otherwise incomprehensible data elements demonstrates how by using these domain-agnostic technologies we made nonstandardized health care data interoperable.

ACCOMPANYING CONTENT

Appendix

Accepted November 21, 2025
Published January 14, 2026

JCO Clin Cancer Inform
10:e2500177

INTRODUCTION

Adolescent and young adult (AYA) cancer is rare and concerns an often-overlooked population, defined as people age 15 to 39 years at primary cancer diagnosis. AYAs are in part characterized by significant differences in tumor type, psychosocial characteristics, and care needs. As a result, AYAs cancer calls for age-specific care that is traditionally unmet by pediatric and adult cancer care.[1]

To improve health care services, research, and outcomes for AYAs, the STRONG AYA Initiative[2] is setting up a federated data infrastructure[3,4] that incorporates both retrospective and prospective AYA data—contributed by several medical centers across Europe. This regional variability allows us to highlight significant challenges in data harmonization across health care systems.

The definitions, format, and terminology of the data in these data sets vary across institutions; this is often for practical and operational purposes. For example, when recording an individual's highest obtained educational level, it is most pragmatic to consult participants in a format that is sensible in their regional setting—as this is the information people

CONTEXT

Key Objective

Collaborative research and data reuse is slowed down by data schema interoperability challenges. Standardized schemas would be ideal but often end up being inflexible with regards to evolving research needs, and universal adoption is not always assured.

Knowledge Generated

Semantic web standards can be used to bridge semantic interoperability issues between already-existing data schema. Without enforcing universal syntactic rigidity, the developed linked-data tooling provides interoperability by using semantic mappings in the metadata.

Relevance *(U. Topaloglu)*

This work addresses a barrier in rare disease research, data interoperability, by enabling seamless integration of heterogeneous datasets across institutions. By applying Semantic Web technologies to the STRONG AYA initiative, the authors facilitate more comprehensive and reusable data infrastructure, which can accelerate clinical insights and improve outcomes for adolescents and young adults with cancer.*

*Relevance section written by *JCO CCI* Associate Editor Umit Topaloglu, PhD.

will know. Such differences can render data incompatible with that collected in other regional settings, if not adequately harmonized.

Therefore, one of the first crucial steps in such large and international collaborations is to adopt or establish certain standards within a data model. This involves two distinct components: standardizing the data schema and standardizing the definitions and terminology. Establishing a standardized schema provides syntactic interoperability; but this still relies on semantic interoperability—standardizing definitions and terminology. These two components constitute interoperability in the broader sense, enabling data integration from diverse sources.[5] For educational level, a straightforward solution would involve transcribing such regionally sensible levels to UNESCO's International Standard Classification of Education,[6] but this can be costly and burdensome. Moreover, even for simple concepts, intrinsic differences in data semantics regularly occur. For example, *biological sex* can be recorded as *male, female*; *male, female, intersex*; and *0, 1, 2*. More complex cases are abundant, and even a seemingly straightforward example like the *time of diagnosis* can quickly reveal operational differences. The definition of the *time of diagnosis* can range from the first illness-related hospital visit to the date of biopsy evaluation or formal diagnosis made by a clinician.

A well-established approach to solve such challenges is the implementation of the F.A.I.R.—findable, accessible, interoperable, and reusable—data guiding principles.[7] This previous work defines how data can be made F.A.I.R. for both humans and machines, while allowing flexibility in terms of multiple coexisting semantic ontologies schema and semantics. In a federated data infrastructure, applying the F.A.I.R. principles has been effective in overcoming interoperability hurdles related to data semantics,[8,9] while bridging pitfalls concerning syntactic interoperability, for example, using an on-read approach.[9] At the same time, the F.A.I.R. principles highlight reusability, thus increasing the understanding of data—and hence transforming it into information—which is a pivotal aspect of the process. To that end, the use of knowledge graphs and semantic web standards are established methods as they are semantically rich and reflect the structure of the data at hand.[8-12] These concepts represent complex information in a graphical format and, therewith, aim to enhance understanding of the data by illustrating relationships between data concepts. These methodologies are however domain agnostic and lack the specificity relevant for STRONG AYA.

The aim of this work was to develop and implement a tailored data model for STRONG AYA that addresses the unique challenges of AYA cancer data harmonization. To achieve this, we developed a data model for STRONG AYA that is aligned with its data collection procedures, simultaneously delivering on the implementation of the interoperable and reusable aspects of the F.A.I.R. data principles. In this work, we elaborate on this effort to illuminate the various processes involved with resolving data incompatibilities in a large health care consortium's federated infrastructure. To reduce the complexity of concepts relevant for AYAs with cancer, we developed the STRONG AYA data model as a knowledge graph. Using this knowledge graph, we then transcribe its contents to the STRONG AYA semantic map which can be used for the necessary mapping that ensures semantic interoperability, while overcoming syntactic

interoperability through an established on-read approach. With the interplay of the knowledge graph and semantic map, we aim to accelerate and facilitate STRONG AYA's goals of improving health care services, research, and outcomes for AYAs—while also setting an example of F.A.I.R. data principles implementation in large-scale consortia.

METHODS

Data Elements

In STRONG AYA, extensive research identified key information to enhance health care services, research, and outcomes for AYAs with cancer. This involved a literature review, qualitative interviews, and a three-round Delphi procedure with AYA cancer stakeholders (AYAs with cancer, caregivers, health professionals, researchers, and policymakers) to determine relevant outcome domains.[13,14]

A Core Outcome Set (COS) was developed from these domains. Subsequently, a set of core measurement instruments and/or items were compiled which best measure the COS. The COS and measurement set were refined to minimize participant burden while retaining essential elements. A list of relevant case-mix variables, identified through a literature review,[15] supplemented the COS to form the final data elements for the STRONG AYA infrastructure, represented in Table 1, excluding all time elements except the initial timestamp. Details on these procedures can be found in their original publications.[13-15]

TABLE 1. Overview of AYA Cancer Relevant Concepts Compiled Through a Literature Review, Qualitative Interviews, and a Three-Round Delphi Procedure With AYA Cancer Stakeholders

Variable	Source
Annual income	PROM
Gender	PROM
Ethnicity	PROM
Occupational status (18 subconcepts)	PROM
Romantic partnership status	PROM
Identifying as disabled	PROM
Registered as disabled	PROM
Educational level	PROM
Age at initial diagnosis	EHR
Year of initial diagnosis	EHR
Biological sex	EHR
Charlson Comorbidity Index (15 subconcepts)	HCPROM
T-stage	HCPROM
N-stage	HCPROM
M-stage	HCPROM
Chemotherapy	EHR
Hormone therapy	EHR
Radiotherapy	EHR
Stem cell therapy	EHR
Surgery	EHR
Targeted therapy	EHR
Bespoke questions (24 subconcepts)	PROM
EORTC QLQ AYA (30 subconcepts)	PROM
EORTC QLQ C30 (30 subconcepts)	PROM
HADS (14 subconcepts)	PROM
Life-threatening infections (3 subconcepts)	PROM
Mental health referral	EHR
CTCAE organ functioning scores (23 sub-concepts)	HCPROM
Cognitive decline	HCPROM
Time to cancer progression (if applicable)	EHR
Number of follow ups till time of cancer progression	EHR
Survival time	EHR
Reason of death	HCPROM

Abbreviations: AYA, adolescents and young adult; CTCAE, Common Terminology Criteria for Adverse Events; EHR, electronic health record; EORTC QLQ, European Organization for Research and Treatment of Cancer Quality of Life Questionnaire; HADS, Hospital Anxiety and Depression Scale; HCPROM, healthcare professional–reported outcome measure; PROM, patient-reported outcome measure.

Data Conversion and Annotation

For multicenter semantic mapping and knowledge graph generation, we used the Resource Description Framework (RDF) data format.[16] RDF is a data representation standard for the basic building block of a graph: the representation of nodes and arcs. This triple format is made up of a subject—predicate—object statement representing node—arc—node, respectively. For instance, *AYA—has column—biological sex.*

As none of the centers collected data as RDF-triples, we used the *Flyover* tool[9,17] to harmonize the data format across centers. *Flyover* converts an arbitrary form of data such as comma separated values, into triples and then stores them in a graph database. For instance, a row—AYA 1—with a value of *female* for *biological_sex* being converted into *AYA 1—has column—biological_sex* and *biological_sex—has value—female* as is illustrated in Figure 1.

Using this triple format, *Flyover* allows us to impose semantics on top of this existing data through a metadata layer—or annotation graph. This means that the *AYA—has column—biological_sex* triple can refer to variables whose names do not necessarily carry semantic significance, which is one of the challenges that was emphasized in the introduction. *Flyover* maintains the original data structure by providing semantic interoperability on-read through this annotation graph.

To make the best use of *Flyover*'s descriptives abilities, we developed a *JSON* semantic map plugin for *Flyover*'s graphical user interface. Using this semantic map, we can directly map variable names as they appear in their original data source, to standardized terminologies. This semantic map could then be used to easily develop the queries that annotate the

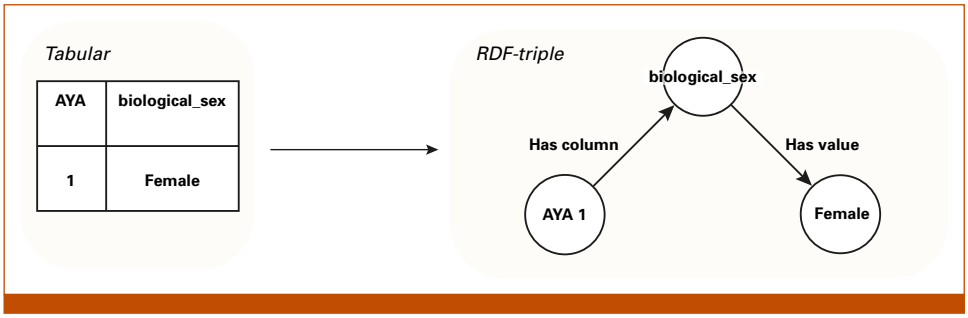

FIG 1. Conversion of tabular biological sex data to RDF-triple format using Flyover. AYA, adolescent and young adult; RDF, Resource Description Framework.

original data sources' names in the metadata layer through *Flyover*'s *Annotation Helper*. This helper parses the variable and values names' along with the mapped standard terms into queries that insert triple statements into the annotation graph.

Knowledge Graphs and Semantic Map

The RDF-model enabled the addition of semantically rich graph structures to the data without modification. Using the list of AYA cancer-relevant concepts, we created a visual knowledge graph to illustrate these data elements in a structured way. We displayed this visual knowledge graph in three substructures, all part of a single graph model: data, data source, and instrument graph.

The graph structure was then reviewed by those who defined the list of AYA cancer relevant concepts elements. After this review, we included the graph structures in the semantic map so that they could be incorporated in the metadata layer.

For the STRONG AYA data elements, we identified relevant standardized terminologies, predominantly leveraging the *National Cancer Institute Thesaurus* (NCIt)[18] for object terms—or classes, and the *Semanticscience Integrated Ontology* (SIO)[19] for properties. Other vocabularies that were used include the *Gender, Sex, and Sexual Orientation Ontology*[20] and *SNOMED CT*.[21] We used custom terms for STRONG AYA's bespoke questions and concepts lacking standardized definitions.

We used *Flyover*'s *Annotation Helper* semantic map format as basis and as overview of all data elements and their standard terms. This global semantic map—without local terms—was published on GitHub to allow for transparent semantic map updates and traceability. The workflow that was used to achieve semantic interoperability is illustrated in Figure 2.

To provide a comprehensive overview of our knowledge graph and semantic map we integrated them into a STRONG AYA Knowledge Representation, semistatic *Hugo*[22] website.

This resource enables continuous review by consortium members. The semantic map section displays variable names, vocabulary reference codes, and associated preferred names and definitions from *BioPortal*.[23] Content is updated quarterly and upon GitHub repository update by extracting semantic map information and fetching vocabulary details via the *BioPortal* REST API. Unfound reference codes are automatically reported as GitHub issues to alert repository owners.

Semantic Mapping Demonstration

As part of STRONG AYA's retrospective data retrieval and to demonstrate our semantic mapping method, we highlight the mapping of the SURVAYA study (ClinicalTrials.gov identifier: NCT05379387), a population-based cross-sectional cohort study of long-term AYA cancer survivors from the Netherlands Cancer Registry. Details can be found in the original study publication.[24] SURVAYA will be integrated into STRONG AYA's infrastructure, but for testing, we used a synthetic data set[25] containing overlapping elements. The SURVAYA study was conducted in accordance with the Declaration of Helsinki and approved by the Netherlands Cancer Institute Institutional Review Board (IRBIRBd18122) on February 6, 2019. The synthetic data set was used with permission from the study's principal investigator and sponsor.

Informed Consent Statement

Informed consent was obtained from all participants involved in the SURVAYA study.

RESULTS

Knowledge Graphs

The knowledge graph in Figure 3 offers a visual and structured representation of AYA cancer-relevant concepts. Specifically, Figure 3 presents the data graph, categorizing data concepts into sociodemographic, clinical, and outcome characteristics using—using SIO's has annotation—to reduce complexity and structure concepts. Data concepts are generally attributes of a given category, using SIO's has attribute. Units for continuous concepts, such as age at diagnosis, overall progression time, and survival, were added in years and days to enhance interpretability, associated via

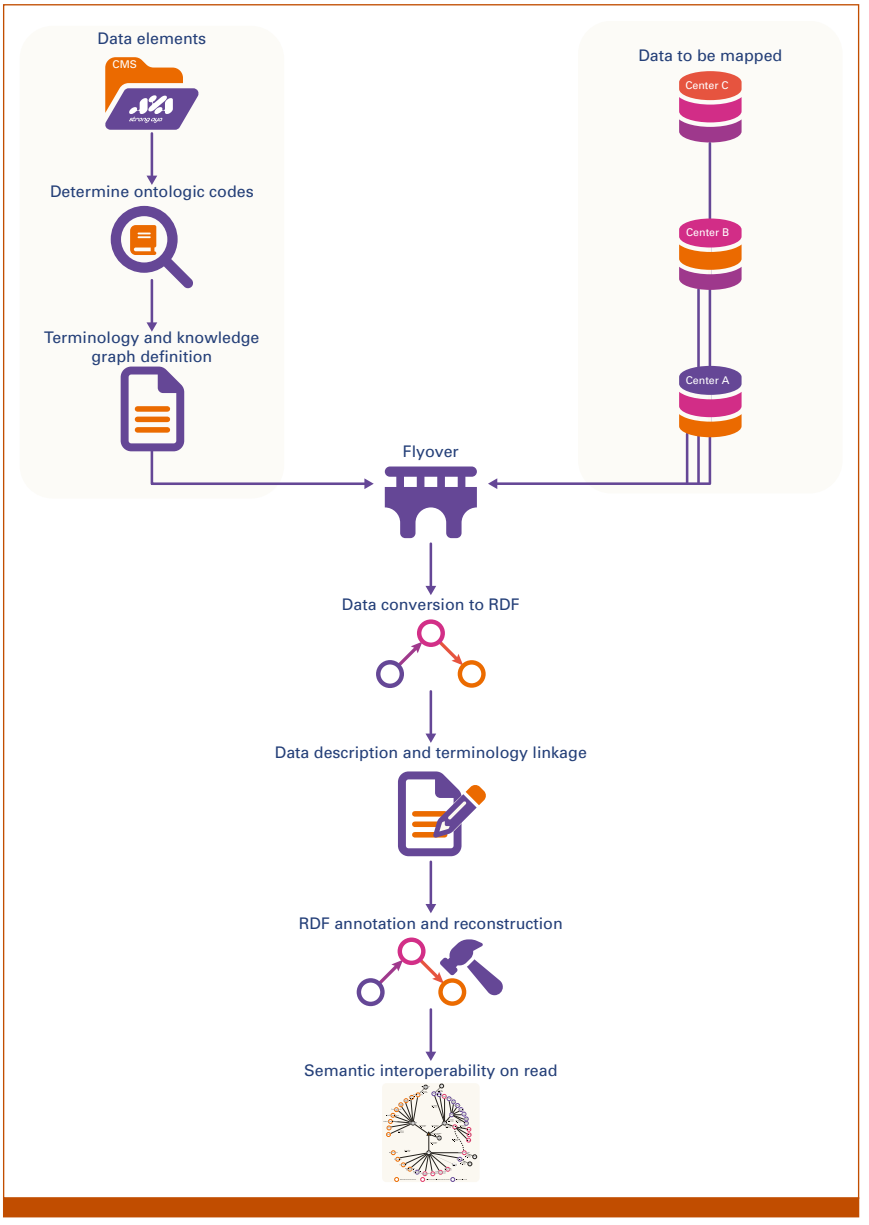

FIG 2. Workflow used to achieve semantic interoperability for STRONG AYA data elements and data-contributing centers. AYA, adolescent and young adult; CMS, Centers for Medicare & Medicaid Services; RDF, Resource Description Framework.

SIO's has unit. Intervariable relationships are limited to neoplasm-associated concepts, including cancer progression time, tumor staging, and localization,—which use SIO's is related to and has property. The AYA's research identifier is directly associated with the AYA using SIO's has unique identifier and is not part of any subcategory. Data sources are color-coded: orange for patient-reported outcome measures (PROMs), pink for health care professional reported outcome measures (HCPROMs), and purple for electronic health records (EHRs).

The data source graphs are displayed in Appendix Figure A1 and describe the data elements' sources using SIO's has property, detailing the distinct properties of PROM, HCPROM, and EHR data elements. Appendix Figure A2 introduces an additional layer to the graph structure by clustering data related to a single concept, exemplified through the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire for AYAs and its specific questions.[26]

Knowledge Representation Webpage

Figure 4 shows an excerpt of the semantic map and the knowledge representation webpage of the concept *biological sex*. The semantic map describes the references to standard vocabularies (in bold orange font) and defines the graph structure (in bold blue font). Concretely, this semantic

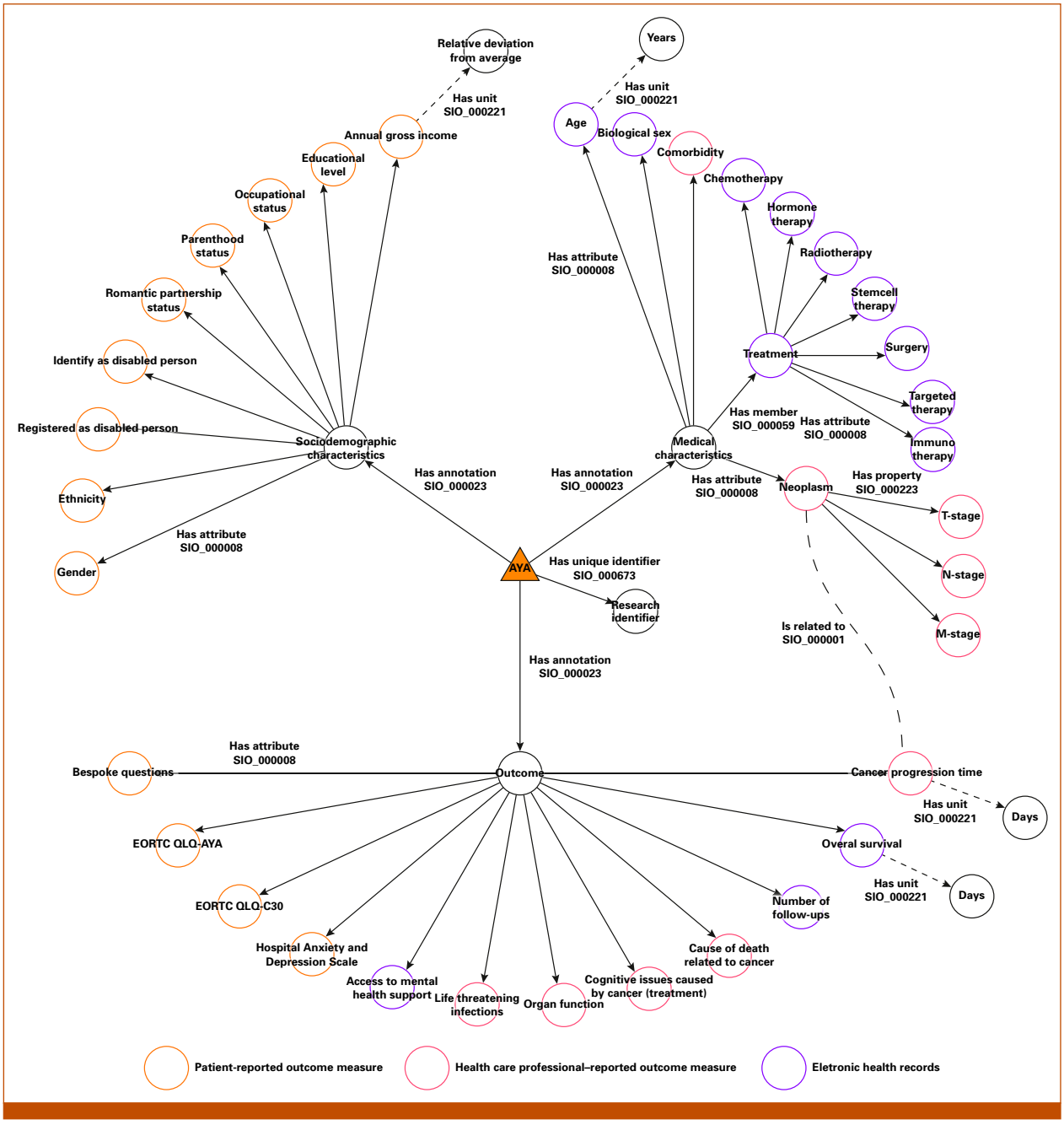

FIG 3. Data graph showcasing the AYA cancer relevant concepts in a more comprehensive way. Information on collection time is not present here and is visible in the underlying instrument graphs. The measurement instrument type or data source per concept is identifiable by the colored outline. Please note that while certain concepts are specifically categorized as Outcome, what constitutes an outcome is study-specific and may also include variables categorized here under Medical characteristics and Sociodemographic characteristics. AYA, adolescent and young adult; EORTC QLQ, European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire; SIO, Semanticscience Integrated Ontology.

mapping annotates our triple statement of *AYA 1—has column—biological_sex* with *AYA 1—sio:SIO_000235—ncit:C18772* and *ncit:C18772—sio:SIO_000008—ncit:C28421*.

The single triple statement is reconstructed to two statements through the schema reconstruction section, which reflects the structure of the previously described knowledge graphs. Our value triple statement of *biological_sex—has value—female* is annotated with *ncit:C28421—has value—ncit:C16576*. Although in the webpage the structure is displayed in a human-readable format by showing what the references in the semantic map—here being the triple statements classes—correspond to.

The complete semantic map[27] and the knowledge representation pages[28] are available on GitHub.

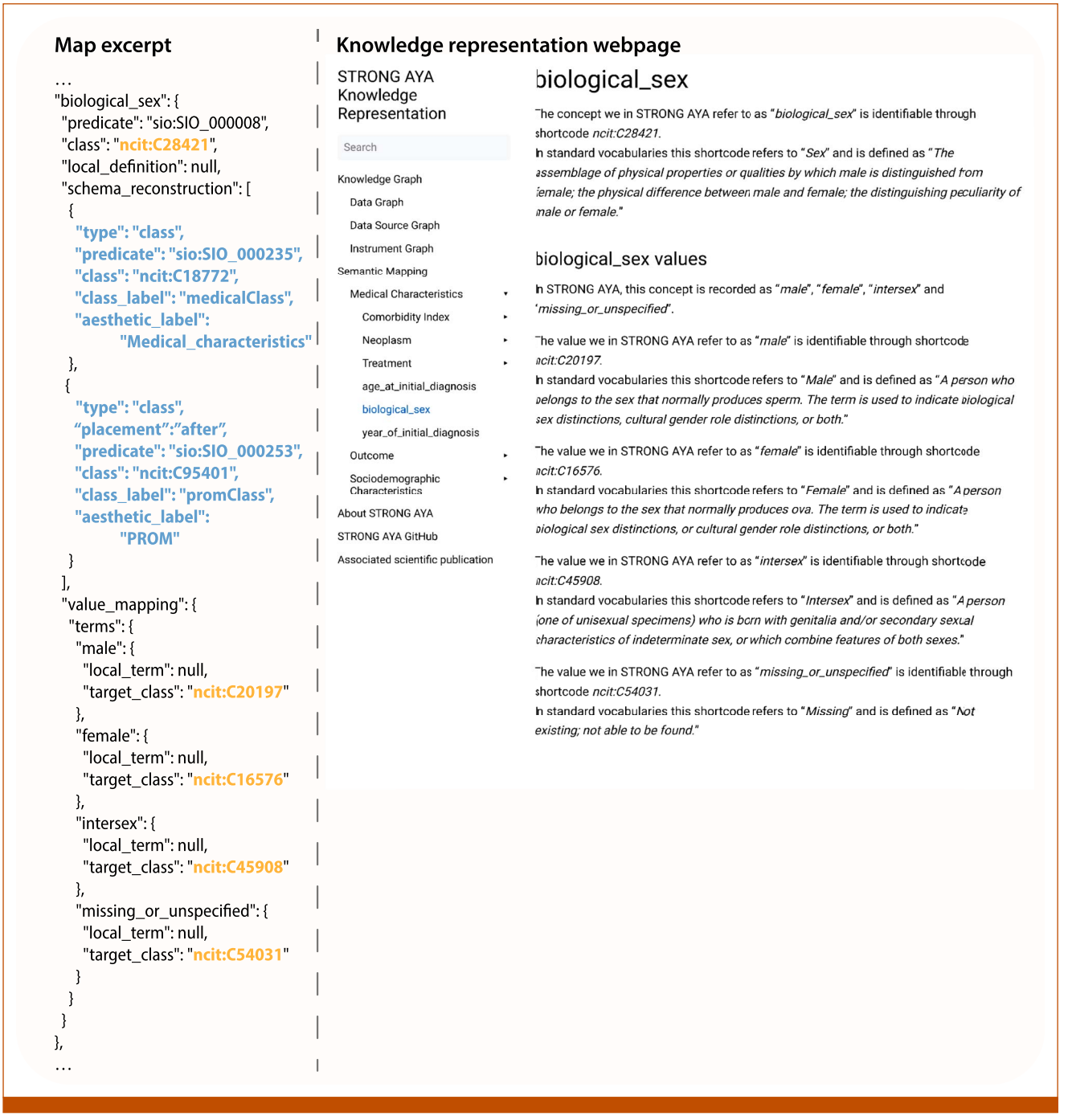

FIG 4. An excerpt of the AYA semantic map and of the AYA cancer knowledge representation. Within the semantic map excerpt the reference to standardized terminologies in bold orange font, and the graph reconstruction in bold blue font. AYA, adolescent and young adult.

Semantic Mapping Demonstration

The interoperability of the annotated RDF-triple SURVAYA data is illustrated in Figure 5, showcasing data accessibility through both local and standard terminology—here exemplified using biological sex. Using *Flyover* and the semantic map, we tested our data harmonization workflow with synthetic SURVAYA data. Initially, RDF-converted biological sex data of a SURVAYA data set would solely be available as *AYA 1—has column—alg_v1b*, but through annotation these data become accessible through the standard semantic mapping of *AYA 1—sio:SIO_000235—ncit: C326200* and *ncit:C326200—sio:SIO_000008—ncit:C28421*. SPARQL[29] was used for data queries. The semantic mapping for SURVAYA is available on GitHub[30] alongside other data sets from multiple international institutions mapped for inclusion in the STRONG AYA ecosystem in the repository's branches. The time taken to process these datasets varied largely and

Local terminology

SPARQL-query

```
PREFIX db: <http://data.local/rdf/ontology/>
PREFIX dbo: <http://um-cds/ontologies/databaseontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?AYA ?biological_sex_value
WHERE {
   ?AYA dbo:has_column ?biological_sex .
   ?biological_sex rdf:type db:survaya.alg_v1b .
   ?biological_sex dbo:has_cell ?biological_sex_cell .
   ?biological_sex_cell dbo:has_value ?biological_sex_value .
}
LIMIT 5
```

Result

AYA	biological_sex_value
0	"1.0"
1	"1.0"
2	"1.0"
3	"1.0"
4	"0.0"

Standard terminology

SPARQL-query

```
PREFIX dbo: <http://um-cds/ontologies/databaseontology/>
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX sio: <http://semanticscience.org/resource/>

SELECT DISTINCT ?AYA ?biological_sex_value
WHERE {
   ?AYA sio:SIO_000235 ?medical_characteristics .
   ?medical_characteristics sio:SIO_000008 ?biological_sex .
   ?biological_sex rdf:type ncit:C28421 .
   ?biological_sex dbo:has_cell ?biological_sex_cell .
   ?biological_sex_cell rdf:type ?biological_sex_value .

   FILTER STRSTARTS(STR(?biological_sex_value),
         "http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#")
   FILTER (!REGEX(STR(ncit:C28421), STR(?biological_sex_value))) .

}
LIMIT 5
```

Result

AYA	biological_sex_value
0	ncit:C16576
1	ncit:C16576
2	ncit:C16576
3	ncit:C16576
4	ncit:C20197

FIG 5. Demonstration of the interoperability of the SURVAYA data set, which through a SPARQL-query is both accessible via its original—and incomprehensible—terminology alg_v1b and standardized terminology ncit:C28421. Both the local terminology and the standard terminology are highlighted in bold font. Shown values are from synthetic SURVAYA data and do not contain information of real participants. AYA, adolescent and young adult; NCIt, National Cancer Institute Thesaurus.

was directly correlated with data set dimensions. Appendix Figure A3 presents the semantic mapping excerpt for SURVAYA's biological sex data, emphasizing the necessity of semantic meaning, as local terminology lacks clarity without it. The full output of our graph database containing synthetic SURVAYA data is available in Appendix Figure A4.

DISCUSSION

The knowledge representation described in this work illustrates how, by making use of RDF-data structures, we can make a large mix of complex, AYA cancer concepts to adhere to the interoperable and reusable aspects of the F.A.I.R. data principles.[7] Our work demonstrates that adhering to these principles allows us to navigate through the difficulties of heterogeneous semantics and inherent differences in data schemas while simultaneously expanding the application of a domain-agnostic standard such as RDF.

The STRONG AYA Knowledge Graph has provided a comprehensive overview of the large number of items collected for this consortium. This knowledge graph was then transcribed into a STRONG AYA data semantic map. In turn, the introduction of this semantic map enables us to map and annotate AYA-specific concepts with standardized terminologies, thereby circumventing the use of project-specific definitions. We showcase how we use the AYA knowledge representation on one of the AYA data sets to be included in STRONG AYA, laying a foundation for other data sets.

In the process of making data more F.A.I.R., there are numerous approaches to address challenges related to data interoperability and enhance understandability. In this work, we used semantic web standards, such as RDF[16] and SPARQL,[29] because of the flexible and nonrigid schema design of the RDF-format. This flexibility allows data-contributing partners to submit data in its original format, with most interoperability work managed by a single coordinating party. Partners can use the *Flyover* tool and Annotation Helper plugin with the STRONG AYA semantic map for most scenarios. The introduced Annotation Helper significantly simplifies this process, providing an easy-to-understand aid verifiable by those without RDF or semantic web knowledge. The annotation helper also enhances RDF's flexibility, as the annotation layer can be reapplied as requirements evolve. For instance, the introduced instrument graph supports simultaneous use of cross-sectional and longitudinal data, but as prospective data collection procedures evolve, the instrument graph must adapt. Additionally, the RDF-format allows for future inclusion of logical reasoning to identify erroneous combinations in the STRONG AYA Knowledge Graph, benefiting data quality assurance and efficient data use.

In comparison with other data models, such as the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM),[31] our approach benefits from *Flyover*'s on-read approach,[9] allowing the STRONG AYA knowledge graph and schema to adapt without modifying the data. Additionally, inherently to semantic web standards and the main mantra "Anyone can say Anything about Anything" we are not limiting ourselves to given terminology standards and data schemas.

By using uniform resource identifiers of terms—which should resolve to their descriptions—these standards are more open to extension by anyone. However, this flexibility means forgoing the tools available for OMOP-CDM, which, despite its rigidity, offers a well-established ecosystem for data integration and analysis. By contrast, RDF provides the flexibility crucial for STRONG AYA's diverse and evolving data structures. However, rather than comparing RDF to OMOP, they should be considered complementary. RDF is a data representation standard predominantly used by big-tech and industry,[32,33] while OMOP is a health care-specific data model. Future work should focus on integrating OMOP and RDF to create a sustainable hybrid solution.

When developing any form of knowledge representation, it is vital that the included concepts are relevant to the overarching subject. In our work, we have based our AYA cancer knowledge representation on an extensive Delphi procedure and literature review.[13,15] This procedure significantly reduced difficulties in defining the relevant concepts—reiterating the importance of such preparatory work. Adherence to this protocol ensured that the true meaning of a term such as *date of diagnosis* was already quite refined and prevented extensive discussions during data model development. Because of this robustness and rigorous adherence to these predefined concepts, we were required to use numerous custom and thus nonstandard terminologies. This is because a substantial number of AYA cancer concepts are not present in any ontology, owing to the bespoke nature of various PROMs. Although this does not hinder interoperability, the lack of standard terms diminishes understandability, as our custom ontology codes have no established semantic significance.

Moreover, although this work advances knowledge representation for AYA cancers, it also highlights areas lacking common definitions. It underscores both the need for extended AYA research and the establishment—and adoption—of standard PROM terminologies.[34]

All in all, we have shown how developing an AYA knowledge representation can navigate the challenging topography of interoperability and understandability in a large AYA cancer consortium. By leveraging existing tools and terminologies, we can increase the adherence of our AYA data pool to the F.A.I.R. data principles while concurrently reducing the burden for our data-contributing partners by using *Flyover*'s integrated *Annotation Helper*.

Although issues of incompatibility and understandability seem addressed, further implementation of F.A.I.R. data

principles will enhance the societal benefits of the data collected in STRONG AYA.

Future work should focus on transcribing our knowledge representation to a machine-findable source that, through appropriate agreements, licenses, and protocols is accessible to individuals currently outside of the consortium—simultaneously addressing the Findable and Accessible attributes in F.A.I.R. To maximize the reusability for a wider audience, it is however a necessity that future work also focuses on the introduction of standardized terminology for AYA cancer—and other fields with notable reliance on PROMs.

AFFILIATIONS

[1]Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, the Netherlands
[2]School of Health Sciences, University of Southampton, Southampton, United Kingdom
[3]Clinical Standards Unit, British Association of Dermatologists, London, United Kingdom
[4]Department of Medical Oncology, Netherlands Cancer Institute, Amsterdam, the Netherlands
[5]Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Centre, Rotterdam, the Netherlands
[6]Department of Public Health, Erasmus MC Cancer Institute, Erasmus University Medical Centre, Rotterdam, the Netherlands
[7]Department of Public Health and Surgical Oncology, Erasmus Medical University Centre, Rotterdam, the Netherlands
[8]Brightlands Institute for Smart Society (BISS), Faculty of Science and Engineering, Maastricht University, Maastricht, the Netherlands
[9]Institute of Molecular Medicine, RWTH Aachen University, Aachen, Germany

CORRESPONDING AUTHOR

Joshi Hogenboom, MSc; e-mail: joshi.hogenboom@maastrichtuniversity.nl.

SUPPORT

Supported in part by the European Union's Horizon 2020 research and innovation program through The STRONG AYA Initiative (Grant agreement ID: 101057482; authors: J.H., V.G., C.C., S.H.M.J., K.W., A.L.A.J.D., A.D., O.H., and L.Y.L.W.), and BETTER project (Grant agreement ID: 101136262; author J.S.), NWO DACIL (KICH1.GZ03.21.023; author: A.L.G.), ERDF DigiONE-I3 (SEP-210898024; author: A.L.G.), ZonMW (author: L.Y.L.W.), Velux Stiftung (author: L.Y.L.W.), NWA-ORC (author: L.Y.L.W.), and an institutional grant by Eli Lilly outside of this work (author: W.T.A.G.).

DATA SHARING STATEMENT

The knowledge graph, semantic map, and knowledge representation website are available on GitHub: https://github.com/STRONGAYA/AYA-cancer-semantic-map, and Zenodo: https://doi.org/10.5281/zenodo.17521899. The Flyover tool is available on GitHub: https://github.com/MaastrichtU-CDS/Flyover, and Zenodo: https://doi.org/10.5281/zenodo.17419800.

AUTHOR CONTRIBUTIONS

Conception and design: Joshi Hogenboom, Varsha Gouthamchand, Andre L.A.J. Dekker, Leonard Y.L. Wee, Johan van Soest, Aiara Lobo Gomes
Financial support: Leonard Y.L. Wee
Administrative support: Leonard Y.L. Wee
Provision of study materials or patients: Silvie H.M. Janssen
Collection and assembly of data: Joshi Hogenboom, Silvie H.M. Janssen, Kirsty Way
Data analysis and interpretation: Joshi Hogenboom, Varsha Gouthamchand, Charlotte Cairns, Silvie H.M. Janssen, Kirsty Way, Winette T.A. van der Graaf, Anne-Sophie Darlington, Olga Husson, Leonard Y.L. Wee, Aiara Lobo Gomes
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).

Andre L.A.J. Dekker
Employment: Medical Data Works B.V
Stock and Other Ownership Interests: Medical Data Works B.V
Honoraria: Janssen-Cilag, Philips Research (Inst), Varian Medical Systems (Inst), IQvia
Research Funding: Varian Medical Systems (Inst), Philips Healthcare (Inst), IQvia (Inst)
Patents, Royalties, Other Intellectual Property: Royalties from Mirada Medical Ltd on a deep learning application in medical imaging (Inst), Royalties from Varian Medical Systems on dosimetry software for radiation oncology (Inst), Royalties from Health Innovation Ventures on Radiomics, eLearning and prediction models in cancer (Inst), Royalties from PXI for software related to small animal irradiation for life sciences research (Inst), Patents held currently -Monitoring respiration based on plethysmographic heart rate signal. US patent 6,702,752. -Apparatus and method for monitoring respiration with a pulse oximeter. US patent 6,709,402. -Monitoring mayer wave effects based on a photoplethysmographic signal. US patent 6,805,673. -Monitoring physiological parameters based on variations in a photoplethysmographic baseline signal. US patent 6,896,661. -Monitoring physiological parameters based on variations in a photoplethysmographic signal. US patent 7,001,337. -Knowledge-Based

Interpretable Predictive Model for Survival Analysis. US Patent 8,078,554 -Dose distribution modeling by region from functional imaging. US Patent 8,812,240 -Systems, methods and devices for analyzing quantitative information obtained from radiological images US Patent 9721340 B2
Expert Testimony: Thoratec
Travel, Accommodations, Expenses: Medical Data Works B.V, IQvia

Winette T.A. van der Graaf
Research Funding: Lilly (Inst)

Anne-Sophie Darlington
Uncompensated Relationships: EORTC Quality of Life Group (Inst)

Johan van Soest
Leadership: Medical Data Works
Stock and Other Ownership Interests: Medical Data Works

Aiara Lobo Gomes
Research Funding: IQvia (Inst), Johnson & Johnson/Janssen (Inst)
Travel, Accommodations, Expenses: IQvia (Inst)

No other potential conflicts of interest were reported.

REFERENCES

1. GBD 2019 Adolescent Young Adult Cancer Collaborators: The global burden of adolescent and young adult cancer in 2019: A systematic analysis for the Global Burden of Disease Study 2019. Lancet Oncol 23:27-52, 2022
2. The STRONG AYA consortium. 2025. https://strongaya.eu/
3. Martin F, Beusekom B, Leurs R, et al: vantage6. 2025
4. Smits D, Van Beusekom B, Martin F, et al: An improved infrastructure for privacy-preserving analysis of patient data. Stud Health Technol Inform 295:144-147, 2022
5. Almeida JR, Silva LB, Bos I, et al: A methodology for cohort harmonisation in multicentre clinical research. Inform Med Unlocked 27:100760, 2021
6. UNESCO IoS: International Standard Classification of Education (ISCED). 2017
7. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al: The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018, 2016
8. Gaudet-Blavignac C, Raisaro JL, Toure V, et al: A national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research within the Swiss personalized health network: Methodological study. JMIR Med Inform 9:e27591, 2021
9. Gouthamchand V, Choudhury A, Hoebers FJP, et al: Making head and neck cancer clinical data findable-accessible-interoperable-reusable to support multi-institutional collaboration and federated learning. BJR Artif Intel 1:ubae005, 2024
10. Sloep M, Kalendralis P, Choudhury A, et al: A knowledge graph representation of baseline characteristics for the Dutch proton therapy research registry. Clin Transl Radiat Oncol 31:93-96, 2021
11. Sachdeva S, Bhalla S: Using knowledge graph structures for semantic interoperability in electronic health records data exchanges, information. 2022
12. Scoarta S, Kucukosmanoglu A, Bindt F, et al: Review: A roadmap to use nonstructured data to discover multitarget cancer therapies. JCO Clin Cancer Inform 10.1200/CCI.22.00096
13. Husson O, Janssen SHM, Reeve BB, et al: Protocol for the development of a Core Outcome Set (COS) for adolescents and young adults (AYAs) with cancer. BMC Cancer 24:126, 2024
14. Darlington AS, Way K, Collaço N, et al: Development of a Core Outcome Set (COS) for adolescents and young adults (AYAs) with cancer. 2025
15. Janssen SHM, van der Graaf WTA, Hurley-Wallace A, et al: Core patient-centered outcomes for adolescents and young adults with cancer: A comprehensive review of the literature from the STRONG-AYA project. Cancers 17:454, 2025
16. Cyganiak R, Wood D, Lanthaler M: RDF 1.1 Concepts and Abstract Syntax, W3C. 2014
17. Gouthamchand V, Hogenboom J, Wee L, et al: Flyover. Zenodo, 2025. 10.5281/zenodo.17419800. https://github.com/MaastrichtU-CDS/Flyover
18. de Coronado S, Remennik L, Elkin PL: National Cancer Institute Thesaurus (NCIt), in Elkin PL (ed) Terminology, Ontology and Their Implementations. Health Informatics. Cham, Switzerland, Springer International Publishing, 2023, pp 395-441
19. Dumontier M, Baker CJ, Baran J, et al: The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semantics 5:14, 2014
20. Kronk CA, Dexheimer JW: Development of the gender, sex, and sexual orientation ontology: Evaluation and workflow. J Am Med Inform Assoc 27:1110-1115, 2020
21. Millar J: The need for a global language—SNOMED CT introduction. Stud Health Technol Inform 225:683-685, 2016
22. Pedersen BE, Francia S, Fok A, et al: https://gohugo.io. 2025
23. Whetzel PL, Noy NF, Shah NH, et al: BioPortal: Enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 39:W541-W545, 2011
24. Vlooswijk C, Poll-Franse LVV, Janssen SHM, et al: Recruiting adolescent and young adult cancer survivors for patient-reported outcome research: Experiences and sample characteristics of the SURVAYA study. Curr Oncol 29:5407-5425, 2022
25. Hogenboom J, Lobo Gomes A, Dekker A, et al: Actionability of synthetic data in a heterogeneous and rare health care demographic: Adolescents and young adults with cancer. JCO Clin Cancer Inform 10.1200/CCI.24.00056
26. Sodergren S, Husson O, Rohde G, et al: Development of an EORTC Quality of Life Questionnaire Specific to Adolescents and Young Adults With Cancer: Measuring What Matters Most to This Unique Patient Group. Cologne, Germany, ISOQOL, 2024, pp S31-S32
27. Hogenboom J: AYA-cancer-semantic-map: STRONG AYA semantic map—Template. Zenodo, 2025. 10.5281/zenodo.17521900. https://github.com/STRONGAYA/AYA-cancer-semantic-map
28. Hogenboom J: AYA-cancer-semantic-map: STRONG AYA semantic map—Web Resource. Zenodo, 2025. 10.5281/zenodo.17521900. https://strongaya.github.io/AYA-cancer-semantic-map
29. Harris S, Seaborne A: SPARQL 1.1 Query Language. 2013, p W3C
30. Hogenboom J: AYA-cancer-semantic-map: STRONG AYA semantic map—SURVAYA mapping. Zenodo, 2025. 10.5281/zenodo.17521900. https://github.com/STRONGAYA/AYA-cancer-semantic-map/tree/dev/retrospective/SURVAYA
31. Hripcsak G, Duke JD, Shah NH, et al: Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers, in Sarkar IN, Georgiou A, de Azevedo Marques PM (eds): MEDINFO 2015: eHealth-enabled Health. Amsterdam, the Netherlands, IOS Press, pp 574-578, 2015
32. He Q: Building the LinkedIn Knowledge Graph. LinkedIn, 2025
33. Singhal A: Introducing the Knowledge Graph: Things, Not Strings. Google, 2012
34. Cella D, Hays RD: A patient reported outcome ontology: Conceptual issues and challenges addressed by the Patient-Reported Outcomes Measurement Information System®(PROMIS®). Patient Relat Outcome Measures 13:189-197, 2022

APPENDIX

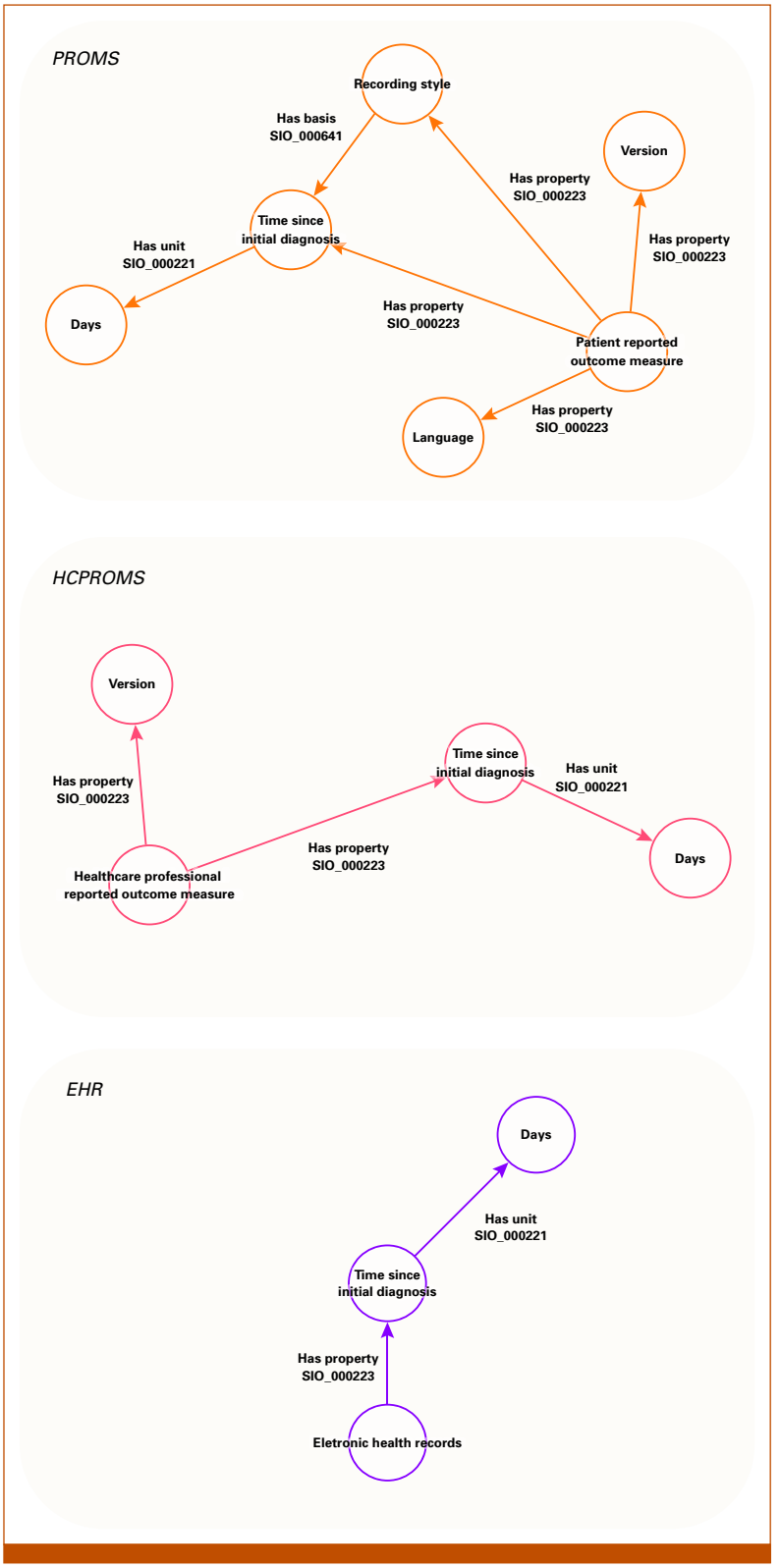

FIG A1. Instrument graphs showcasing the underlying structures of the various measurement instruments—or data sources—in the AYA cancer knowledge representation. Each source type had distinct properties, but all included a variable for time elapsed since diagnosis, derived from the (continued on following page)

FIG A1. (Continued). recording timestamp, with a unit of days via SIO's has unit property. PROMs and HCPROMs included a version property, with PROMs also featuring a language property. Each data source was linked to its data element using SIO's has source property. HCPROM, health care professional–reported outcome measure; PROM, patient-reported outcome measure; SIO, Semanticscience Integrated Ontology.

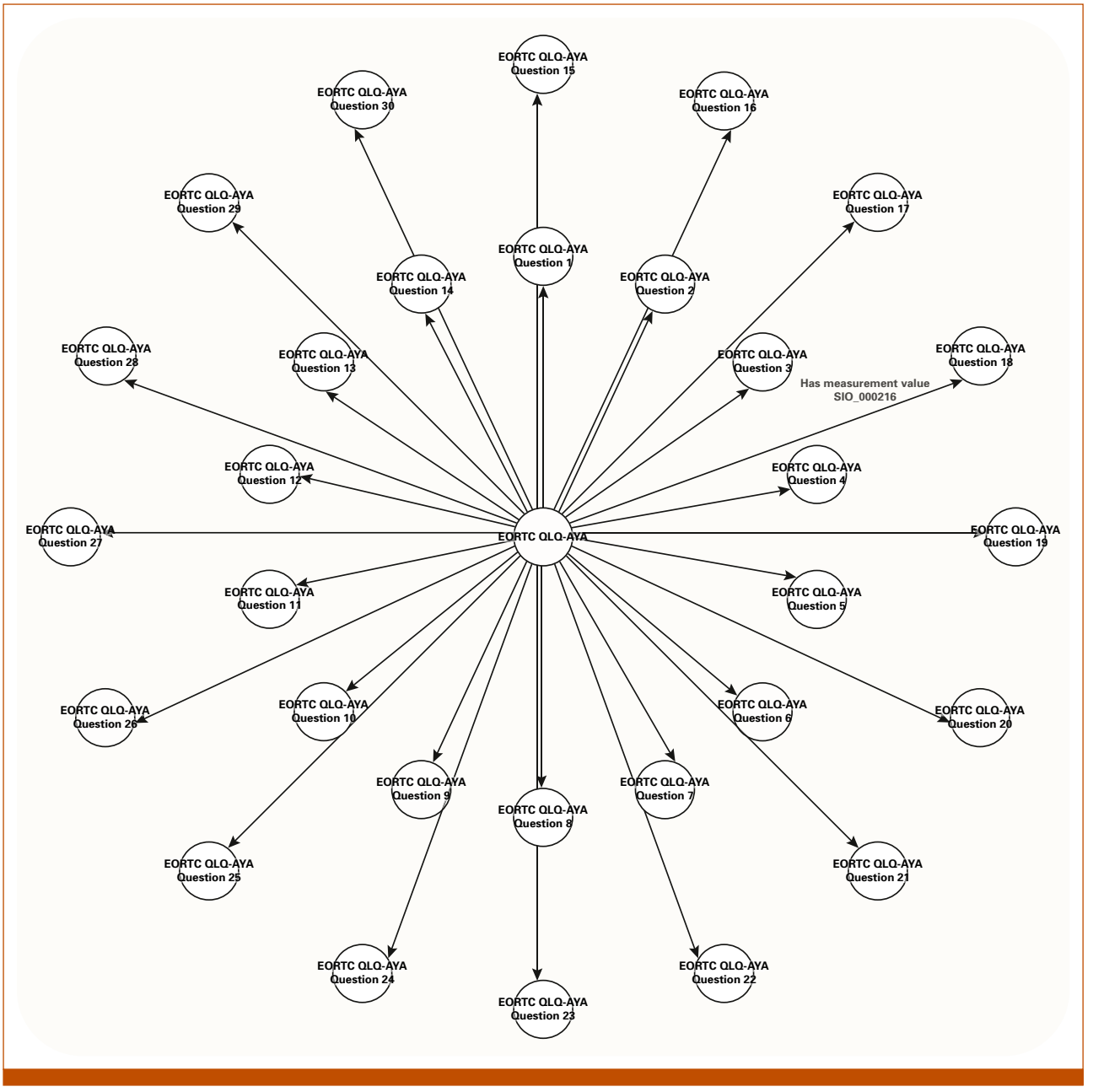

FIG A2. Instrument graph illustrating the underlying measurements of a measurement instrument with multiple subconcepts, such as here exemplified using the EORTC QLQ-AYA patient-reported outcome measure. Generally, instrument graphs comprise the main concept with its subconcepts as measurement values—using SIO's has measurement value. EORTC QLQ-AYA, European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Adolescent and Young Adults; SIO, Semanticscience Integrated Ontology.

Mapping excerpt biological sex

```
…
"biological_sex": {
 "predicate": "sio:SIO_000008",
 "class": "ncit:C28421",
 "local_definition": "alg_v1b",
 "schema_reconstruction": [
  {
   "type": "class",
   "predicate": "sio:SIO_000235",
   "class": "ncit:C18772",
   "class_label": "medicalClass",
   "aesthetic_label":
          "Medical_characteristics"
  },
  {
   "type": "class",
   "placement":"after",
   "predicate": "sio:SIO_000253",
   "class": "ncit:C95401",
   "class_label": "promClass",
   "aesthetic_label":
          "PROM"
  }
 ],
 "value_mapping": {
  "terms": {
   "male": {
    "local_term": "0.0",
    "target_class": "ncit:C20197"
   },
   "female": {
    "local_term": "1.0",
    "target_class": "ncit:C16576"
   },
   "intersex": {
    "local_term": null,
    "target_class": "ncit:C45908"
   },
   "missing_or_unspecified": {
    "local_term": null,
    "target_class": "ncit:C54031"
   }
  }
 }
},
…
```

FIG A3. Demonstration of the mapping for the synthetic SURVAYA data set, with local terminology of biological sex in bold font. NCIt, National Cancer Institute Thesaurus.

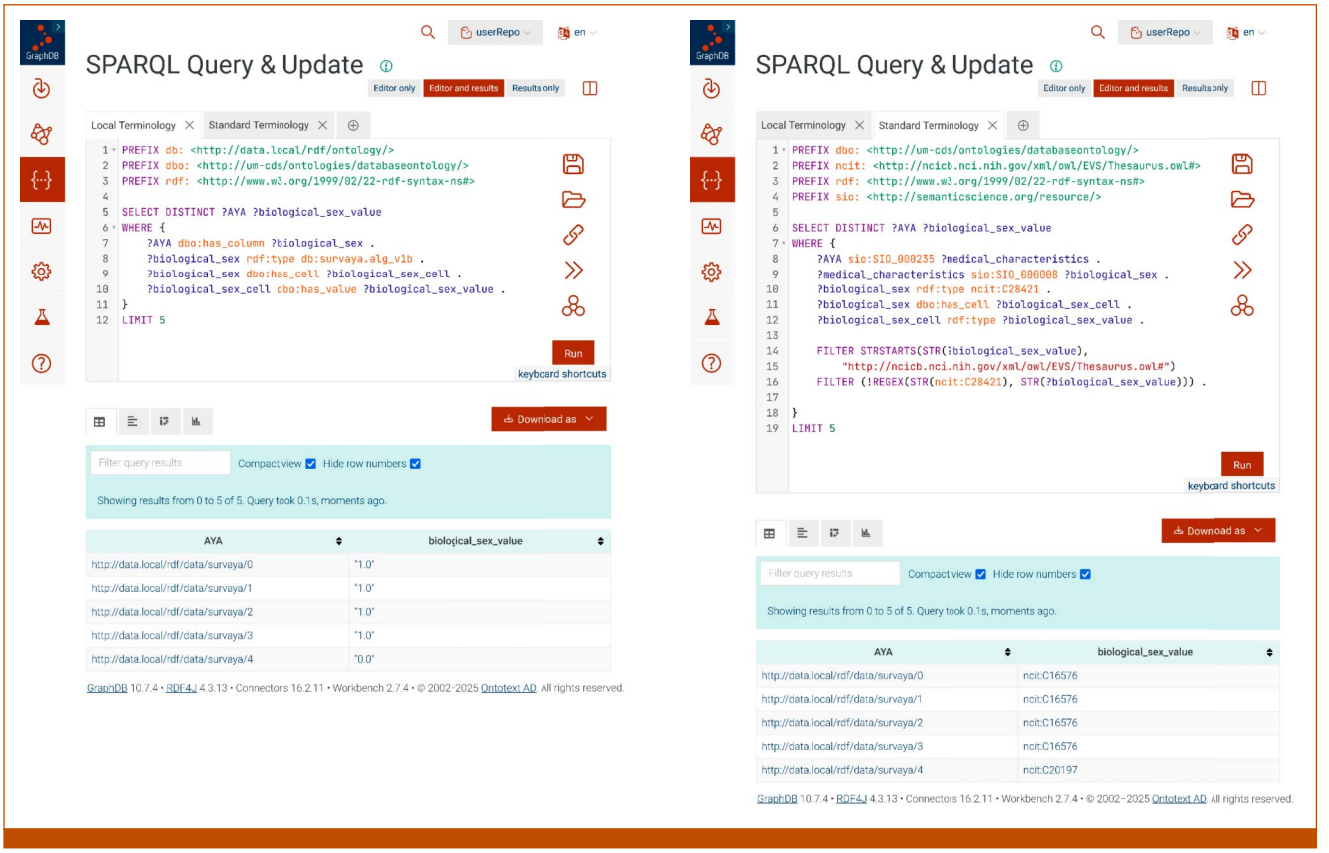

FIG A4. Graph database output demonstrating the interoperability of our synthetic SURVAYA data.