Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The SPHN Federated Clinical Routine Dataset: A Metadata-driven Assessment of Clinical Data Coverage and Terminology Usage

Vasundra Touré

vasundra.toure@sib.swiss

SIB Swiss Institute of Bioinformatics

Jan Armida

SIB Swiss Institute of Bioinformatics

Deepak Unni

SIB Swiss Institute of Bioinformatics

Philip Krauss

Accenture AG

Harald Witte

SIB Swiss Institute of Bioinformatics

Davide Chiarugi

SIB Swiss Institute of Bioinformatics

Andrea Brites Marto

SIB Swiss Institute of Bioinformatics

Julia Mauer

Swiss Academy of Medical Science

Thomas R. Geiger

Swiss Academy of Medical Science

Henning Beywl

Inselspital, Bern University Hospital

Marouan Borja

University Hospital of Geneva

Marc Daverat

University Hospital of Geneva

Olga Endrich

Inselspital, Bern University Hospital

Dominique Furrer

Inselspital, Bern University Hospital

Mathias Gassner

University Children's Hospital Zurich
Yves Jaggi
University Hospital Lausanne
Matthias Joos
University Hospital Zurich
Katie Kalt
University Hospital Zurich
Janshah Veettuvalappil Ikbal
Inselspital, Bern University Hospital
Helena Peić Tukuljac
University Children's Hospital Zurich
Gaëlle Vuaridel-Thurre
University Hospital Lausanne
Karen Triep
Inselspital, Bern University Hospital
Solange Zoergiebel
University Hospital Lausanne
Sabine Österle
SIB Swiss Institute of Bioinformatics

Research Article

Keywords: Federated datasets, Metadata, Knowledge Graph, RDF, Terminology Usage, Real World Data

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11756329/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The SPHN Federated Clinical Routine Dataset: A Metadata-driven Assessment of Clinical Data Coverage and Terminology Usage

Vasundra Touré[1,*,†], Jan Armida[1,†], Deepak Unni[1], Philip Krauss[2], Harald Witte[1], Davide Chiarugi[1], Andrea Brites Marto[1], Julia Mauer[3], Thomas R. Geiger[3], Henning Beywl[4], Marouan Borja[5], Marc Daverat[5], Olga Endrich[4], Dominique Furrer[4], Mathias Gassner[6], Yves Jaggi[8], Matthias Joos[7], Katie Kalt[7], Janshah Veettuvalappil Ikbal[4], Helena Peić Tukuljac[6], Gaëlle Vuaridel-Thurre[8], Karen Triep[4], Solange Zoergiebel[8], Sabine Österle[1]

[1] *Swiss Personalized Health Network, SIB Swiss Institute of Bioinformatics, Basel, Switzerland*
[2] *Accenture AG, Basel, Switzerland*
[3] *Swiss Personalized Health Network, Swiss Academy of Medical Science, Bern, Switzerland*
[4] *Inselspital, Bern University Hospital, Bern, Switzerland*
[5] *University Hospital of Geneva, Geneva, Switzerland*
[6] *University Children's Hospital Zurich, Zurich, Switzerland*
[7] *University Hospital Zurich, Zurich, Switzerland*
[8] *University Hospital Lausanne, Lausanne, Switzerland*
* Corresponding author: vasundra.toure@sib.swiss
[†] These authors contributed equally

Abstract

Background: Over the past nine years, the Swiss Personalized Health Network (SPHN) has developed a national federated framework enabling semantically interoperable health-related data, with a primary focus on hospital clinical routine data. The framework supports FAIR (Findable, Accessible, Interoperable, and Reusable)

data principles while preserving patient privacy by avoiding the centralization of patient-level information and enabling data sharing only upon request.

Methods: We conducted a metadata-based analysis of the SPHN Federated Clinical Routine Dataset (FedData) using summary statistics available in the SPHN Metadata Catalog. Metadata was retrieved through the catalog's SPARQL endpoint and analysed using SPARQL queries and R scripts. The analysis focused on the dataset size, SPHN concept coverage, and terminology usage.

Results: As of October 2026, the SPHN FedData comprises more than 959,000 patients who provided broad consent, covering the period from 2018 to present. Across the participating hospitals, the infrastructure holds over 16.5 billion (10^9) RDF triples mapped to 141 SPHN semantic concepts including demographics, diagnoses, procedures, medications, laboratory results, vital signs, clinical scores, allergies, microbiology, intensive care data, oncology, and biological samples.

Conclusion: The SPHN Federated Clinical Routine Dataset demonstrates how a federated, semantically interoperable infrastructure can support large-scale health data reuse. By combining local data stewardship with centralized metadata-based discoverability, the SPHN framework provides a robust foundation for research applications and contributes to the broader goals of national and European health data ecosystems.

Keywords

Federated datasets, Metadata, Knowledge Graph, RDF, Terminology Usage, Real World Data

Background

Hospitals generate large amounts of clinical data as part of routine patient care, yet leveraging this data for biomedical research remains challenging. A central barrier is semantic interoperability: heterogeneous data sources may differ in their data models, terminologies, formats, and standards, making it difficult to interpret and combine information consistently across institutions. In Switzerland, linguistic diversity further contributes to data heterogeneity as clinical information may be documented in different languages such as German, French or Italian. Without a common semantic representation, integrating multi-center clinical data can require substantial manual curation, mapping, and normalization, limiting the efficient reuse of routine clinical data for secondary research. These challenges are particularly relevant to collaborative research, where combining data from multiple institutions can increase statistical power and population diversity. However, differences in the representation and availability of clinical information across institutions may hinder the identification and characterization of suitable datasets before analysis can begin. Providing researchers with standardized and discoverable descriptions of available data can therefore facilitate the assessment of multi-center data resources while reducing the need for preliminary manual investigation.

The Swiss Personalized Health Network (SPHN) seeks to address these challenges within Switzerland's national health ecosystem [1]. SPHN promotes standardized data representation and integration, fostering FAIR (Findable, Accessible, Interoperable, Reusable) [2] clinical data across healthcare institutions. A key component of this effort is the SPHN RDF Schema [3], a formal, semantic data model for representing clinical concepts consistently across sites [4]. It leverages well-established international standards such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED

CT, [5]), Logical Observation Identifiers Names and Codes (LOINC, [6]), International Statistical Classification of Diseases and Related Health Problems, 10th Revision, German Modification (ICD-10-GM, [7]), International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3, [8,9]), Orphanet Rare Disease ontology (ORDO, [10]), Digital Imaging and Communications in Medicine (DICOM, [11]), and Anatomical Therapeutic Chemical (ATC, [12]) classification system to enhance interoperability. Swiss hospitals actively contribute to this national endeavor to construct and deliver clinical routine data according to SPHN specifications, ensuring a standardized and semantically rich dataset suitable for research use.

Centralizing patient-level clinical routine data across multiple hospitals would raise significant privacy and legal challenges because such data contains sensitive personal health information subject to data protection regulations and institutional governance policies. The SPHN approach therefore retains patient-level data within the participating institutions. Each hospital generates a standardized local extract of clinical routine data according to the SPHN semantic specifications, and the metadata about these local datasets collectively constitutes the “SPHN Federated Clinical Routine Dataset” (SPHN FedData). Rather than centrally storing patient-level data, the SPHN FedData provides only aggregated descriptive metadata about the contributing datasets, while patient-level data remains securely stored within each hospital. The centrally managed metadata enables researchers to explore data distributions without revealing sensitive information. This supports cross-institution discoverability of data while upholding privacy and governance regulations. Access to patient-level data for specific research projects remains subject to the data access procedures and governance requirements of the respective institution.

This manuscript extends a previous conference paper [13] with an updated quantitative characterization based on a subsequent data release and revised metadata. Following the Semantic Web Applications for Healthcare and Life Sciences (SWAT4HCLS) conference publication, participating hospitals performed a new data load based on an updated release of the SPHN RDF Schema 2026.1, and data quality issues identified during the previous loading cycle were corrected. One hospital that contributed data to the conference version (University Hospital Basel) did not contribute data to the present dataset; consequently, the current analysis covers five hospitals instead of six. All reported numbers, cohort statistics, and quantitative characterizations presented in this manuscript have been revised and supersede those of the earlier conference version. Building on this updated resource, we characterize the scope and content of the SPHN FedData through metadata-derived statistics, offering a more comprehensive and up-to-date view of the resource.

Methods

Data specification

The SPHN FedData contains clinical routine data from patients available in the Clinical Data Platforms of the participating German- and French-speaking hospitals, currently University Children's Hospital Zurich (KISPI), University Hospital Bern (INSEL), University Hospital Geneva (HUG), University Hospital Lausanne (CHUV) and University Hospital Zurich (USZ). Patients were included in the dataset only if they had provided broad consent (referred to as general consent in Switzerland) and had at least one administrative case from 2018 onward (Figure 1). The broad consent allows for

secondary use of patient-level health data for research projects that are not yet defined at the time of consent. In addition, retrospective data may be included, particularly for patients who meet the inclusion criteria and for whom relevant historical data are available. For these patients, data available in the respective clinical data platform and mapped to the SPHN RDF Schema 2026.1 [14] are represented in the SPHN FedData. Individual time series measurements for vital signs are not represented directly as RDF instances, instead metadata describing these data types are represented through the concept "Time Series Data File". The participating hospitals are referred to as H1, H2, H3, H4, and H5 in all quantitative analysis throughout the remainder of this paper.

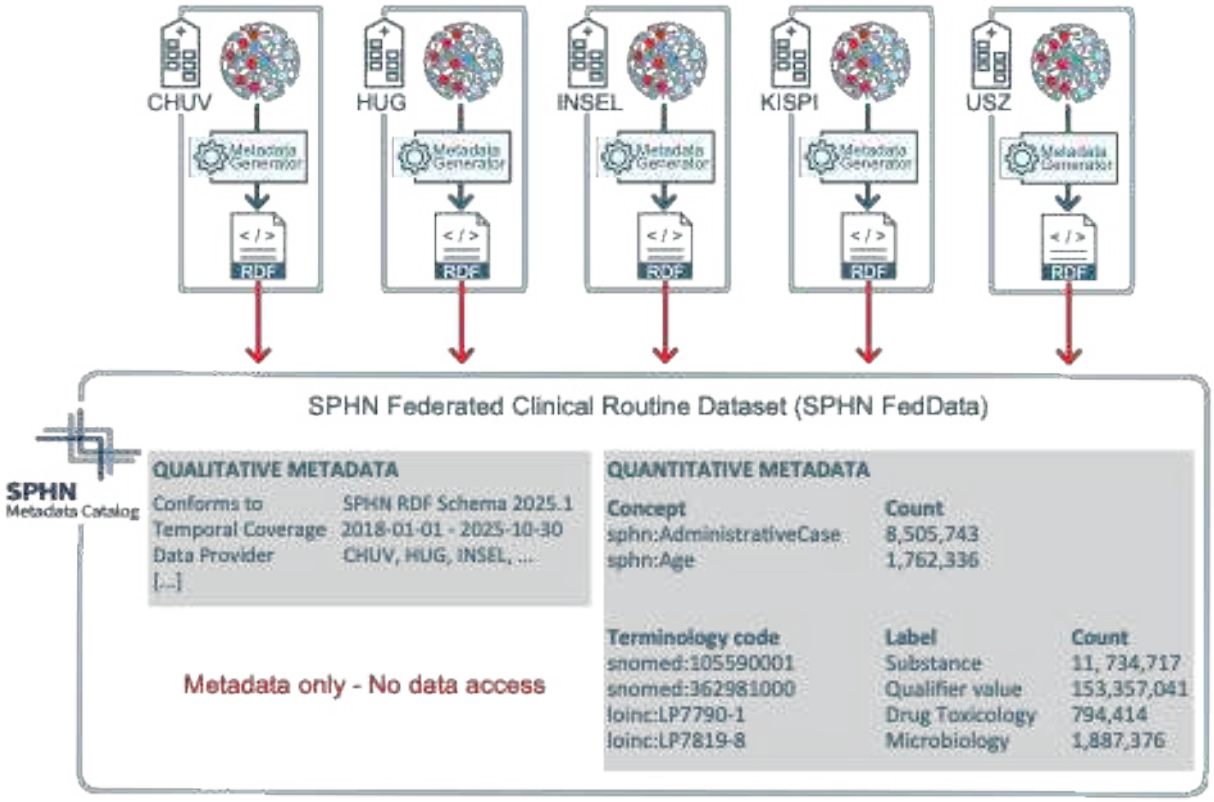


Figure 1. Overview of SPHN FedData metadata extraction and aggregation process. At each participating hospital, the SPHN Metadata Generator tool [15] is applied locally to the knowledge graph representing the hospital's routine clinical data to extract metadata. The extracted metadata are then parsed, transformed and aggregated in the SPHN Metadata

Catalog [16], which provides the qualitative and quantitative metadata describing the SPHN FedData. The SPHN Metadata Catalog is publicly accessible and enables exploration of the federated dataset, while the underlying hospital knowledge graphs containing clinical data remain restricted to the respective institutions.

Data generation and storage

In SPHN, clinical information is modeled as a network of well-defined concepts described by standardized attributes ('composedOfs' relationships) and linked through explicit semantic relationships formalized in the SPHN RDF Schema [17]. Schema-compliant representations are generated through local Extract, Transform, Load (ETL) pipelines using the SPHN Connector [18], a standardized, dockerized component that enables institutions to transform relational or JSON sources into RDF graph structures aligned with the SPHN Semantic Interoperability Framework (henceforth referred to as SPHN framework). The SPHN Connector accepts data from the clinical data platform, optionally performs de-identification (this step can also be carried out using hospital-specific tools) according to defined rules, transforms the extracted data into RDF, validates the resulting RDF graphs against SPHN-specific Shapes Constraint Language (SHACL) rules to ensure compliance with the SPHN RDF Schema, and transforms the resulting dataset into semantically interoperable, SPHN-compliant RDF data. Coding of the data is performed by the hospitals as part of these local pipelines. The resulting data is securely stored in a local Virtuoso [19] instance at each hospital. Each individual dataset and its metadata are regularly updated, and data from patients who have revoked their consent are promptly removed by the hospitals. The data analyzed for this paper adheres to the SPHN RDF Schema 2026.1 [3] and represents a snapshot of the metadata as of 2026-10-23.

Metadata extraction

Qualitative metadata follows the SPHN Metadata Catalog Schema [16,20] and describes general characteristics of the SPHN FedData, including dataset titles and descriptions, temporal coverage, and access conditions provided by the participating hospitals. These metadata provide contextual information about the datasets. Quantitative metadata is generated locally by each participating hospital using the SPHN Metadata Generator (Figure 1). These metadata comprise counts for instances of concepts and predicates, as well as descriptive statistics characterizing concept-level data availability, including the average, minimum and maximum number of concept instances and overall count of concept instances per patient.

In addition, the quantitative metadata includes the frequency of individual codes for the 18 terminologies provided in RDF in SPHN [21]. In the context of the SPHN FedData, 9 of them are delivered by the hospitals: ATC, Swiss Classification of Hospital Operations (CHOP, [22]), DICOM, ICD-10-GM, ICD-O-3, LOINC, ORDO, SNOMED CT, Unified Code for Units of Measure (UCUM, [23]). For codes from coding systems that are not part of the set of SPHN terminologies but may still be used within hospitals, the SPHN RDF Schema allows these elements to be structured using the SPHN "Code" concept. These are stored in text fields, making them semi-structured. Both the qualitative and quantitative metadata were converted into RDF by a customized script and made available on the SPHN Metadata Catalog [16] (Figure 1). The catalog provides a SPARQL endpoint to facilitate metadata querying [24], for instance, it provides frequencies of individual codes when code-level statistics are available.

Metadata analysis

The metadata of the SPHN FedData was analyzed using the SPHN Metadata Catalog's SPARQL endpoint with SPARQL queries and R scripts: the only exception is the part assessing the age distribution of patients. This information is not available from the metadata and was collected from the hospitals directly.

A report of the analysis is available in R Markdown [25] and a snapshot of the metadata is available in Zenodo [26]. In this analysis, we distinguish between 'core' and 'supporting' concepts. Core concepts are entities linked directly to the Subject Pseudo Identifier concept, which represents the patient. Supporting concepts (e.g., codes, quantities, values) are typically reused across different contexts. Rather than linking to the patient directly, they are embedded within core concepts and provide metadata for them. For example, the value 'Body Height' is modeled as a supporting concept and is reused within the core concept 'Body Height Measurement'.

The 2026.1 release of the SPHN RDF Schema (hereafter referred to as '2026.1') comprises 208 SPHN concepts including 82 core concepts. This count excludes the abstract non-instantiable concepts 'SPHN Concept' and 'Measurement'. To describe the domain coverage of the SPHN FedData, concepts were grouped into thematic domains according to their clinical or administrative context, or intended use in the SPHN semantics. The concepts were grouped into ten domains: Administrative, Clinical condition, Demographics, Laboratory, Measurement, Medical procedure, Medication and treatment, Oncology, Imaging, and Provenance (Table 1). The assignment of SPHN concepts to domains was performed manually based on their intended clinical or administrative purpose. Supporting concepts were assigned according to their primary intended use in relation to the corresponding core concepts.

Concepts reused across multiple domains were classified separately as “Cross-domain concepts”. These include ‘Code’, ‘Quantity’, ‘Unit’, ‘Body Site’, ‘Laterality’, ‘Medical Device’, and ‘Intent’. This separation was necessary since the available metadata do not provide sufficient contextual information to determine the domain in which individual instances of these concepts are used. Keeping them separate therefore avoids double-counting instances across domains.

Finally, in this analysis, a concept was considered delivered by a hospital if at least one instance of that concept was present in the corresponding hospital.

[Location of Table 1]

Data quality

Data quality is addressed through a multi-layered approach that distinguishes between quality of the SPHN RDF representation of the data and the quality of the underlying clinical data. Participating hospitals are responsible for the quality of the clinical data extracted from their source systems and provided to the SPHN Connector. The SPHN framework provides automated validation of the resulting data against the SPHN RDF Schema using SHACL, allowing for the checking of structural and semantic constraints such as data types, cardinalities, and terminology bindings. In addition, statistical and consistency checks, implemented through SPARQL queries, can be used to characterize the delivered data and identify potential irregularities, for example by examining the number and distribution of instances, values, codes, or dates.

These mechanisms assess the conformance of the data to the SPHN representation and facilitate the detection of technical and semantic inconsistencies. However, they

do not by themselves establish the correctness or clinical validity of the underlying source data; for example, they cannot determine whether a diagnosis was coded correctly at the source or whether a laboratory value is clinically plausible or within an appropriate clinical range. Such aspects remain dependent on the hospitals' source-level data quality processes and, where relevant, additional clinical or project-specific validation.

Data quality is further monitored iteratively across successive data loads. Based on the metadata available from the delivered datasets, potential anomalies or inconsistencies are identified and communicated to the respective hospitals. The hospitals review the findings against their source systems and mapping processes, allowing observed deviations to be classified as data or mapping issues, expected characteristics of the source data, or artefacts of the data transformation process. Where necessary and possible, corrective actions are implemented in subsequent loads. Conversely, findings may also be documented as expected when they reflect known characteristics or limitations of the source systems.

Results

This section summarizes the main findings regarding the SPHN FedData, focusing on its (i) cohort characteristics, (ii) evolution between 2025 and 2026 (iii) SPHN concepts coverage, and (iv) terminology coverage across participating hospitals. The results illustrate the scale and diversity of the federated datasets, providing insights into the progress towards data standardization achieved within SPHN.

Cohort characteristics

As of 2026, the SPHN FedData includes data from five hospitals, covering 959,978 patients who provided broad consent and had an administrative case between 2018 and 2026 (older data was included if available for these patients). The demographic composition of the cohort is depicted in Figure 2. The cohort is evenly distributed by administrative sex, with 51.2% female patients and 48.8% male patients. The largest age groups are between 60 and 79 years, while the smallest groups of patients are 90 years (or older), or between 20-29 years. In terms of data volume, the SPHN FedData comprises a total of 1.63×10^{10} RDF triples across all local RDF databases combined (Table 2). Note that a total of 73,617 patients are known to be deceased in the shown statistics in Figure 2.

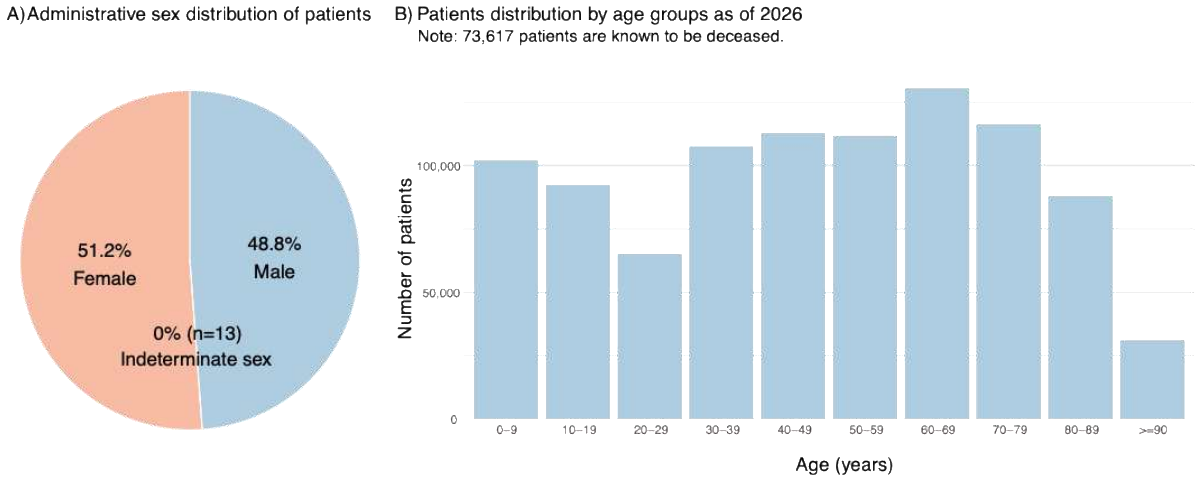


Figure 2. Distribution of patients by sex and age group as of 2026. (A) Patient distribution by administrative sex, including male, female, and indeterminate sex, indicated as SNOMED CT codes in the metadata. For each category, the exact count and percentage is given. (B) Distribution of patients across age groups, with the number of patients shown above each bar.

Table 2. General key statistics for individual hospital datasets. The table reflects the number of Subject Pseudo Identifier (i.e., "Patients"), triples, SPHN concepts, and core concepts available at each individual hospital (columns H1-H5). Aggregate total numbers across all hospitals (H1-H5) are also provided in column "Total". SPHN concepts include all concepts defined in 2026.1 while the core concepts represent a subset which only includes concepts directly linked to a Subject Pseudo Identifier.

	H1	H2	H3	H4	H5	**Total**
Patients	99608	131778	275756	118658	334178	**959978**
Triples [x10^9]	2.2	4.3	5.3	0.7	4.0	**16.5**
SPHN concepts	92	126	96	78	101	**141**
SPHN core concepts	44	62	46	36	44	**68**

SPHN FedData evolution between 2025.1 and 2026.1

Between the 2025.1 version (published in [13]) and the 2026.1 version (analyzed in this paper), the SPHN FedData showed an overall increase in the dataset volume and concept coverage, despite a reduction in the number of participating hospitals from six to five (Table 3). The total number of SPHN concepts delivered increased from 125 to 141, with the number of core concepts rising from 60 to 68 (+13%). The number of core concepts shared across all hospitals also increased from 18 to 26. However, this change should be interpreted in the context of the altered composition of the participating hospitals and does not necessarily indicate improved harmonization across sites. The total number of triples increased from 1.2x10^{10} RDF triples to

294 $1.6x10^{10}$. The most pronounced increases are observed in the terminologies with the
295 number of distinct codes delivered rising from 41,140 to 64,800 (+ 57.5%). The number
296 of instances of codes increased from 750 million to 804 million instances (+ 114%).
297
298 *[Location of Table 3]*
299
300 Since 2025, the SPHN FedData went through iterative processes for improving the
301 quality of the data. The main types of quality improvements include:

302 • *Correction of terminology mappings:* Review and correction of mappings where
303 concepts were associated with inappropriate or overly specific/general
304 terminology codes, including LOINC, SNOMED CT, and ATC.
305 • *Removal or correction of obsolete data:* Identification of concepts or terminology
306 codes that were no longer present or were incorrectly delivered, followed by their
307 removal or correction in subsequent data loads.
308 • *Improvement of concept contextualization:* Identification of concepts appearing
309 in potentially inappropriate contexts, such as vital-sign measurements
310 represented as laboratory results, prompting review and, where appropriate,
311 adjustment of their representation or mapping.
312 • *Resolution of inconsistencies in data distributions:* Investigation of unexpected
313 patient-to-instance ratios, unusually high numbers of instances for individual
314 patients, or discrepancies between patient counts and coded instances.
315 • *Clarification of source-specific mappings:* Documentation or refinement of
316 mappings where the source systems provide only generic information and
317 therefore cannot support a more specific standardized concept.
318

SPHN concepts coverage

The SPHN FedData comprises 141 out of the 208 defined SPHN concepts, including 68 of the 82 core concepts (Table 2). These concepts cover a broad range of domains including demographics, diagnoses, procedures, medications, laboratory (including microbiology and samples), measurements (e.g., vital signs), and allergies.

Concept availability varies across hospitals as only 26 of the 68 are common to all five hospitals (Figure 3). These include: Administrative Case, Administrative Sex, Age, Allergy, Assessment Event, Billed Diagnosis, Billed Procedure, Birth, Blood Pressure Measurement, Body Height Measurement, Body Mass Index, Body Temperature Measurement, Body Weight Measurement, Circumference Measurement, Consent, Death, Drug Administration Event, Electrocardiographic Procedure, Gestational Age At Birth, Healthcare Encounter, Heart Rate Measurement, Lab Test Event, Microorganism Identification Lab Test Event, Nationality, Respiratory Rate Measurement and Sample. This variability reflects differences in local data capture and availability, which may affect data comparability and feasibility of multi-site analyses across specific domains.

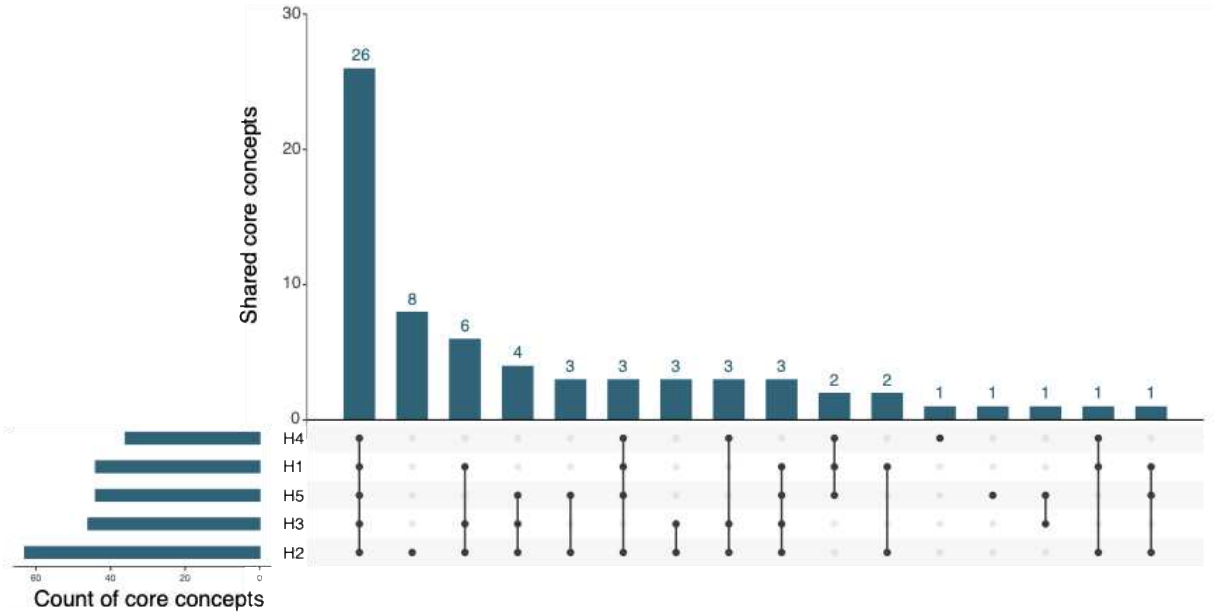


Figure 3. Overlap of shared core concepts across hospitals (H1-H5). Vertical bars indicate the number of core concepts shared among hospitals while horizontal bars indicate the total number of core concepts delivered in each hospital. The dotted lines indicate the number of hospitals and specify which hospitals provide the same core concepts.

Beyond concept coverage, the distribution of concept instances across domains reveals a substantial difference in the volume of the data (Figure 4). For instance, laboratory and measurement domains account for the largest domains instantiated with a total of approximately 966 and 932 million instances, respectively. Clinical condition and administrative domains contribute to intermediate volumes, while provenance, imaging, demographics, medical procedures and oncology account for considerably fewer instances.

Cross-domain concepts are concepts that can be reused across multiple domains. They are presented as a separate category to avoid double-counting instances across domains, as the available metadata do not provide sufficient contextual information to determine the domain in which each instance of these concepts is used.

The contribution of core and supporting concepts also varies across domains, with supporting concepts accounting for a larger share in most high-volume domains, whereas core concepts predominate in the administrative and demographic domains.

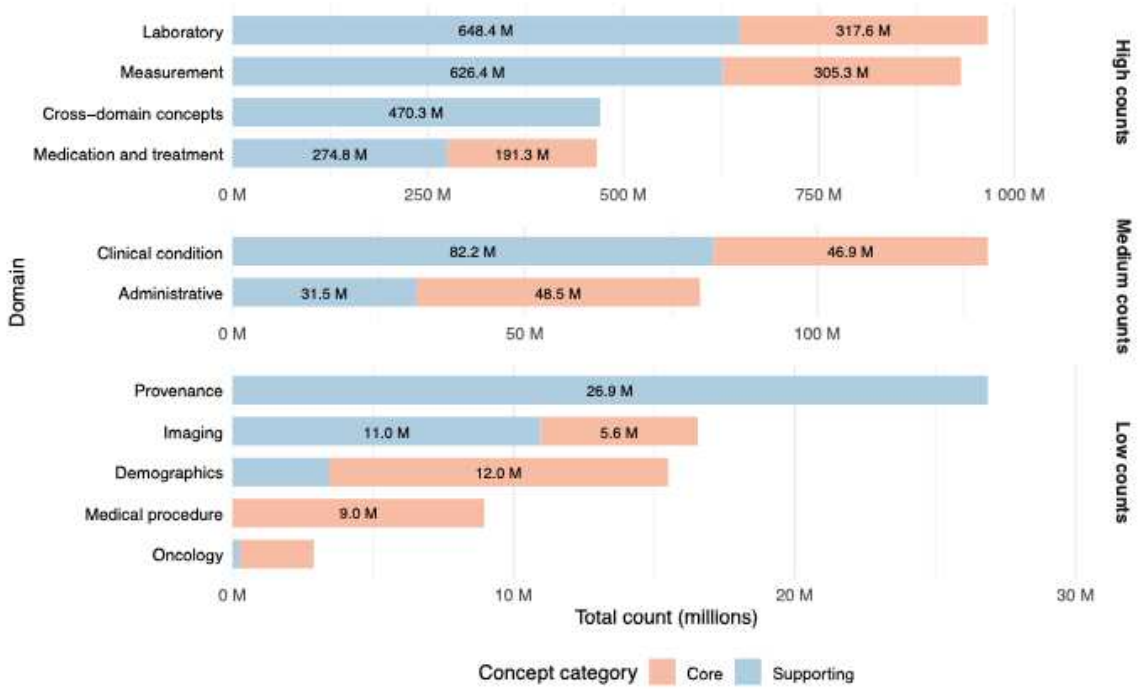


Figure 4. Distribution of instances of core and supporting concepts across domains. Horizontal stacked bar plots showing the total number of concept instances in each domain, stratified by concept category (core and supporting). Domains are grouped into three panels (very high, medium, and low counts) for better readability. Each panel has an independent x-axis scale to facilitate comparison across domains with significantly different counts. Values are expressed in millions (M). The 'cross-domain concepts' reflect concepts that can be reused across multiple domains.

Terminology coverage

In SPHN, a set of 18 terminologies are supported within the SPHN RDF Schema, enabling hospitals to represent data using standardized and uniquely identifiable

codes. Across all participating sites, approximately 1.6 billion instances of terminology codes were identified, with 64,800 distinct codes, originating from 9 out of the 18 available terminologies. The remaining terminologies were not delivered, and thus not represented, in this studied dataset.

Figure 5A and 5B provide detailed information on the terminologies for which codes have been delivered: LOINC, SNOMED CT, ICD-10-GM, CHOP, ATC, DICOM, ICD-O-3, ORDO, and UCUM.

In terms of the breadth of the terminology used, excluding one outlier hospital with a particularly high number of SNOMED CT codes, ICD-10-GM and CHOP show the broadest coverage across hospitals in terms of distinct codes. In contrast, in terms of volume of data, LOINC and SNOMED CT exhibit higher instantiated codes. CHOP and ICD-10-GM were already used for coding billing information, prior to the SPHN initiative, whereas LOINC and SNOMED CT were largely adopted as standards since the beginning of SPHN.

A cross-hospital comparison of terminologies usage, visualized in Figure 5B, reveals substantial overlap for CHOP and ICD-10-GM codes. This overlap reflects the terminologies' established and harmonized use in administrative and billing workflows. In contrast, all other terminologies, including SNOMED CT and LOINC exhibit site-specific variability. Some terminologies are not delivered in some hospitals (e.g., ATC, DICOM, ICD-O-3 and ORDO).

Among the used terminologies, we observe the following:

- SNOMED CT serves as a cross-domain ontology for detailed clinical concepts and phenotype definitions, contributing to approximately 40% of all distinct

terminology codes. SNOMED CT codes are highly instantiated by all hospitals.

- ICD-10-GM and CHOP are used primarily for billing diagnoses and procedures, comprising approximately 21% and 24% of all distinct terminology codes, respectively. Their coding exhibits considerably higher cross-hospital consistency, indicating a mature and stable coding practice with well-defined usage due to harmonized billing processes.
- LOINC is used for laboratory-related concepts (e.g., lab tests, microbiology) and reflects 5.7% of distinct terminology codes. Similar to SNOMED CT, a high amount of LOINC coded instances has been observed.
- ORDO identifies and classifies rare disease codes and reflects 5% of distinct terminology codes.
- ATC codes describe administered or prescribed active substances or substance groups within medications, contributing to approximately 3% of distinct terminology codes.
- ICD-O-3 is used to classify diseases of oncology and reflect 1% of distinct terminology codes.
- UCUM codes are used to describe units and contribute to 0.3% of distinct terminology codes.
- DICOM used for medical imaging concepts reflect 0.02% of distinct terminology codes, and is currently only delivered by one hospital (Figure 5B).

Collectively, these findings indicate that administrative terminologies (ICD-10-GM and CHOP) are well standardized across institutions, yet clinical and laboratory terminologies (LOINC and SNOMED CT) and domain specialized terminologies

(ORDO, ICD-O-3 and DICOM) remain heterogeneous or less used and represent priority areas for ongoing interoperability enhancement within the SPHN framework.

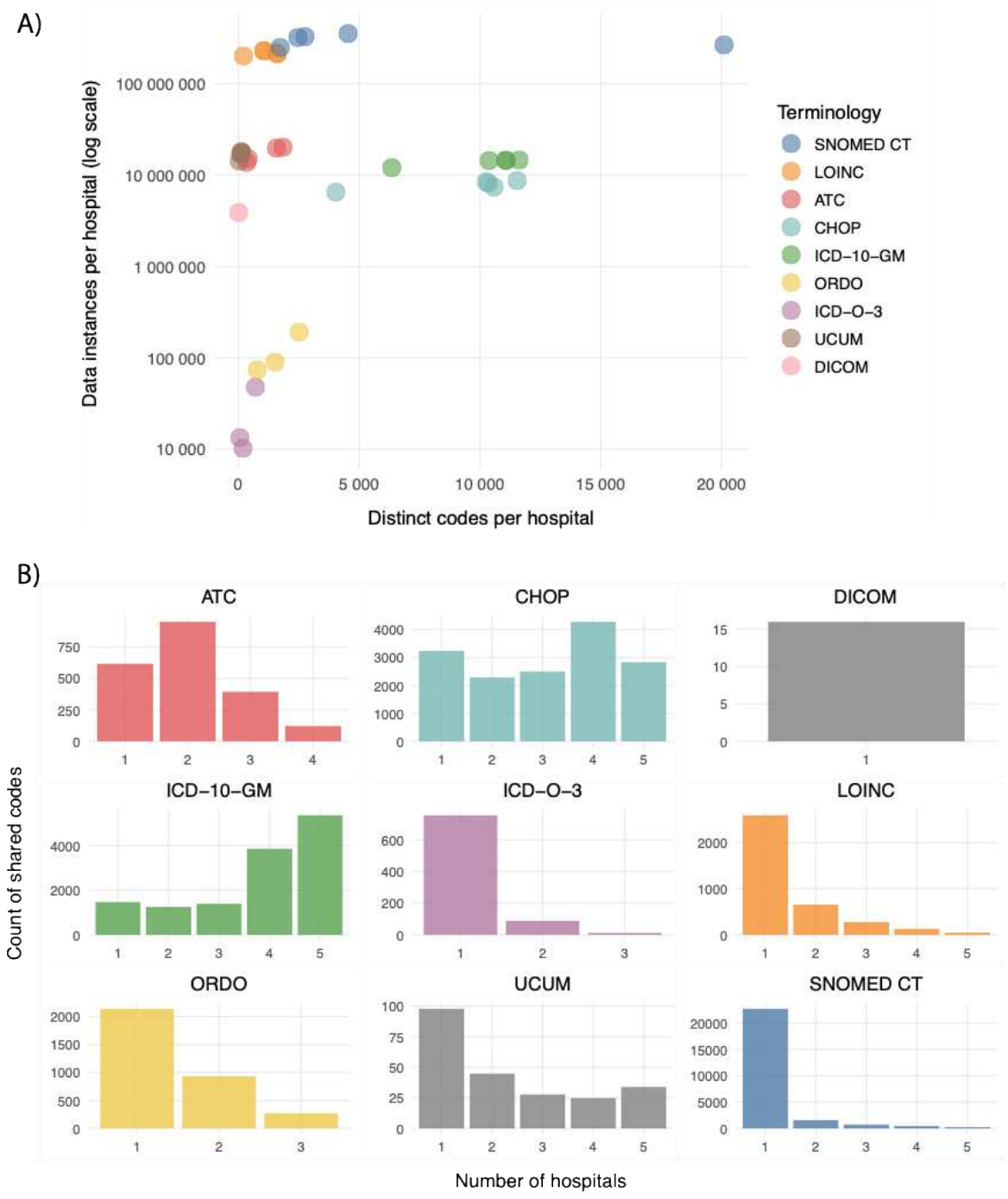


Figure 5. A) Scatter plot of the number of distinct terminology codes per hospital (x-axis, continuous scale) against the number of data instances per hospital (y-axis,

logarithmic scale). Each point represents a hospital, with colors indicating the terminology. The plot illustrates variation across hospitals in both code diversity and the volume of coded data instances. **B) Distribution of shared codes across hospitals by terminology.** The bar plots show the number of shared codes identified in one to five hospitals for each terminology. Note that not all terminologies were provided by all participating hospitals (DICOM n=1, ICD-O-3 n=3, ORDO n=3, and ATC n=4). Each panel represents a terminology with independent y-axis scales.

Discussion

The SPHN FedData builds on the principles of semantic interoperability by representing clinical routine data according to a common semantic framework and aligning selected data elements with international terminologies, participating hospitals can describe their data in a more uniform and machine-readable manner. This provides an important basis for national-scale data-driven projects, cohort discovery, and cross-hospital feasibility assessments. With that, the ability to identify and interpret available data consistently can directly affect the reliability and efficiency of subsequent research activities. Data are validated against the SPHN RDF Schema, during the data generation process, ensuring that the resulting representations are semantically compliant with the defined schema. However, semantic compliance does not guarantee the accuracy, completeness, or clinical appropriateness of the underlying data values; these aspects remain dependent on the source systems and local data-generation and quality-control processes.

The SPHN FedData should be understood as a continuously evolving resource rather than a static dataset. Comparison with the dataset underlying the original conference

publication illustrates substantial progress in both the volume and scope of the data delivered by participating hospitals. The earlier 2025.1 release contained 125 delivered concepts out of 168 concepts available for implementation, whereas the current 2026.1 dataset contains 141 delivered concepts out of 208 available concepts. Thus, despite the reduction from six to five contributing hospitals, the current dataset contains approximately 150,000 additional patients and 4.2 billion additional RDF triples compared with the previous load. This increase is not attributed solely to the larger patient population: the continued implementation of additional SPHN concepts contributes substantially to the amount of represented clinical information. This is particularly evident for imaging-related concepts introduced in the SPHN RDF Schema 2025.2, which are currently delivered by only one participating hospital yet account for approximately 30 million concept instances. Consequently, the addition of even a single high-volume concept domain can result in a substantial increase in the number of patient-level instances and corresponding RDF triples. The results therefore provide evidence of continued maturation of the SPHN data infrastructure at the level of the clinical content represented in the SPHN FedData resource.

The same development is visible in the use of standardized terminologies. Our previous study focused on the adoption of SNOMED CT and LOINC across participating hospitals and dived into their hierarchical structures [27]. Here, we extend this analysis to a broader range of terminologies, remaining at the count level, providing an assessment of terminology coverage and usage in the updated 2026.1 dataset. An increase in the total number of terminology instances is partly expected because of the larger patient population and the increasing number of delivered concepts. More informative, however, is the increase in the number of distinct codes represented in

the datasets. This reflects an increasing richness of the mappings and coded information provided by the hospitals. A particularly pronounced example is SNOMED CT, for which the number of distinct codes increased substantially in the current release. This increase is mainly associated with a strategic decision to integrate data from a new clinical information system at one participating hospital, where SNOMED CT is used as a core reference terminology. This illustrates how the semantic richness of the federated resource can increase as hospitals connect additional source systems and progressively map locally available information to the SPHN framework.

One of the main strengths of the SPHN FedData lies in its metadata-driven discoverability. The SPHN Metadata Catalog provides researchers with complementary means of exploring the resource through descriptive metadata, concept-level statistics, and terminology usage through code-level counts without exposing sensitive personal information. The ability to inspect these statistics for individual hospital datasets and for the federated dataset allows researchers to assess the scope and potential suitability of the resource before requesting access to individual-level data. Additionally, the method is scalable: as new hospitals are onboarded and data volume increases, researchers can continue to explore the dataset using metadata alone, thereby reducing the need for repeated feasibility requests to individual hospitals. Furthermore, it supports equitable access to health data, ensuring that researchers across institutions can benefit from interoperable datasets to generate scientific knowledge for the broader public good. At the same time, the importance of hospitals' data governance structures and data-access procedures remains fully acknowledged. The development of the SPHN Data Explorer is expected to extend these capabilities by supporting cross-hospital feasibility queries

and simple exploratory analyses on aggregated hospital clinical routine data. The system will allow researchers to investigate aggregated data and perform exploratory analyses such as value distributions and compute basic statistics for defined cohorts while preserving patient privacy through privacy-enhancing techniques. Such capabilities complement the metadata-oriented exploration provided by the SPHN Metadata Catalog by allowing researchers to move from describing what data are available to assessing whether those data are potentially suitable for a specific research question. These frameworks enable hospitals to safeguard patient rights and ensure that any access to health data, whether at the patient level or not, is conducted responsibly and in full compliance with ethical and legal regulations. By streamlining access pathways and reducing administrative complexity, such structures also help avoid opportunity costs that arise when valuable data remain underused. In turn, more efficient data utilization prevents delays in research, innovation, and decision-making that could otherwise limit potential health and societal benefits.

Several technical challenges emerged during implementation. The quantitative metadata analysis highlights important differences in how clinical information is represented. The use of terminologies does not necessarily imply semantic equivalence at the level of individual codes. For example, a drug substance may be represented using the ATC code B01AC06, acetylsalicylic acid, whereas another hospital may represent aspirin using the SNOMED CT concept 387458008, Aspirin (substance). Although both representations concern aspirin-related medication information, they encode different aspects of meaning. The ATC representation places acetylsalicylic acid within a therapeutic classification and therefore reflects its pharmacological or utilization context, whereas the SNOMED CT representation

522 describes the clinical substance itself. Consequently, SNOMED CT can provide more
523 precise representation of ingredient-level semantics, while ATC is particularly useful
524 for therapeutic classification, utilization analyses, and reporting.

525 More generally, differences in local coding practices contribute to semantic
526 heterogeneity. The same clinical information may be represented using different
527 terminologies, different codes within a terminology, or different levels of granularity.
528 For example, laboratory tests represented using LOINC may differ in the inclusion of
529 method information or may be associated with different units, while SNOMED CT
530 concepts may be selected at different levels of specificity. This contrasts with
531 terminologies whose use is more strongly aligned across institutions, such as CHOP
532 or ICD-10-GM, and with terminologies for which implementation choices are more
533 dependent on local systems and mapping practices, such as LOINC, SNOMED CT,
534 and ATC. The federated approach preserves local flexibility and allows hospitals to
535 progressively improve their mappings, but it also creates a continuing need for
536 curation, mapping maintenance, and cross-site semantic alignment. The longitudinal
537 increase in terminology coverage observed in the current dataset should therefore be
538 regarded not only as an increase in data volume but also as evidence of an ongoing
539 process of semantic harmonization. Despite the wide use of terminologies, the SPHN
540 FedData also contains more than 60 million instances of the generic “SPHN Code”
541 concept. However, the available metadata do not provide sufficient information to
542 characterize all of these instances in detail. Some may correspond to codes from
543 coding systems that are used locally in a specific hospital but are not currently
544 represented as standardized SPHN terminology resources, effectively providing
545 structured but only partial semantically characterized information. Other instances may
546 originate from coding systems which are currently not provided in RDF by SPHN but

are used in hospitals to code data. This includes, for example, the Global Trade Item Numbers (GTIN, [28]), or unique identifiers for medicinal products.

This highlights both a limitation of the current metadata representation and an opportunity for future semantic enrichment: mapping additional locally used identifiers and coding systems to standardized representations could further increase the interoperability and analytical value of the resource.

Another important limitation concerns data quality. The SPHN RDF Schema and associated validation mechanisms ensure that data conform to the expected structural and semantic model, but they do not by themselves establish whether the underlying clinical information is complete, accurate, timely, or clinically appropriate. Continued quality-control activities at the hospital level remain essential, particularly when additional source systems are integrated or when new concepts and terminologies are introduced.

Conclusions

The SPHN Federated Clinical Routine Dataset represents a major milestone in showcasing the clinical routine data available for research across Swiss institutions. The metadata-driven data discovery approach enables researchers to assess data availability and coverage without initiating lengthy (legal and contractual) procedures to access this information. This supports evidence-based feasibility assessment and helps align subsequent data access requests with research requirements and available data. The SPHN Data Explorer, currently under development, is expected to complement the existing metadata-based discovery capabilities by supporting privacy-

preserving cross-hospital feasibility queries and exploratory analyses on aggregated data.

The longitudinal analysis demonstrates the continued expansion of the resource in terms of patient coverage, clinical concepts represented, and terminology usage, despite changes in the composition of participating institutions. At the same time, differences in concept availability and terminology usage across hospitals highlight persistent challenges in achieving consistent semantic representation and cross-institutional comparability.

Future developments, including the SPHN Data Explorer and the integration of additional healthcare institutions and clinical domains, are expected to further support privacy-preserving feasibility assessment and research-oriented data discovery. In parallel, ongoing harmonization efforts and iterative feedback loops will progressively improve data quality and semantic consistency. Together, these efforts provide a foundation for sustainable, interoperable, and responsible secondary use of clinical routine data in Switzerland.

In conclusion, the SPHN Federated Clinical Routine Dataset demonstrates the feasibility of operating a national-scale federated semantic infrastructure for clinical research. The accompanying SPHN Metadata Catalog not only supports dataset discovery but also provides a scalable mechanism for monitoring interoperability, concept coverage, and terminology adoption across participating institutions while preserving local governance of patient-level data.

References

1. Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. 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 2021 Jun 24;9(6):e27591. doi: 10.2196/27591

2. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, Van Der Lei J, Van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data Nature Publishing Groups; 2016 Mar 15;3(1):1–9. PMID:26978244

3. The SPHN RDF Schema 2026.1. Available from: https://www.biomedit.ch/rdf/sphn-schema/sphn/2026/1 [accessed Oct 30, 2026]

4. Touré V, Krauss P, Gnodtke K, Buchhorn J, Unni D, Horki P, Raisaro JL, Kalt K, Teixeira D, Crameri K, Österle S. FAIRification of health-related data using semantic web technologies in the Swiss Personalized Health Network. Sci Data 2023 Mar 10;10(1):127. doi: 10.1038/s41597-023-02028-y

5. Benson Tim and Grieve G. SNOMED CT. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR Cham: Springer International Publishing; 2016. p. 155–172. doi: 10.1007/978-3-319-30370-3_9ISBN:978-3-319-30370-3

6. LOINC - Logical Observation Identifiers Names and Codes. 2025. Available from: https://loinc.org/ [accessed Nov 17, 2025]

7. ICD-10-GM - International Statistical Classification of Diseases and Related Health Problems, 10th revision, German Modification. 2025. Available from: https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html [accessed Nov 17, 2025]

8. Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, Whelan S. International Classification of Diseases for Oncology (ICD-O). 3rd ed. Geneva: World Health Organization; 2000. Available from: https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology ISBN:9241545348

9. Organization WH. International Classification of Diseases for Oncology (ICD-O) - Third Edition, First Revision (ICD-O-3.1) and Second Revision (ICD-O-3.2). Geneva: World Health Organization; 2019. Available from: http://www.iacr.com.fr/index.php?option=com_content&view=article&id=108:icd-o-3&catid=100:icd-0-3&Itemid=577

10. Vasant D, Chanas L, Malone J, Hanauer M, Olry A, Jupp S, Robinson PN, Parkinson H, Rath A. Ordo: an ontology connecting rare disease, epidemiology and genetic data. Proceedings of ISMB researchgate. net; 2014. p. 1–4.

11. DICOM Standards Committee. Digital Imaging and Communications in Medicine (DICOM). Rosslyn, VA, USA: National Electrical Manufacturers Association; 2026. Available from: https://www.dicomstandard.org/

12. ATC - The Anatomical Therapeutic Chemical . 2025. Available from: https://atcddd.fhi.no/ [accessed Nov 17, 2025]

13. Armida J, Touré V, Krauss P, Unni D, Witte H, Chiarugi D, Marto AB, Mauer J, Geiger T, Beywl H. Semantic interoperability at national scale: the SPHN federated clinical routine dataset. Res Sq 2025;10. doi: https://doi.org/10.21203/rs.3.rs-8250886/v2

14. The SPHN RDF Schema - Release 2025.1. 2025. Available from: https://www.biomedit.ch/rdf/sphn-schema/sphn/2025/1 [accessed Nov 18, 2025]

15. SPHN Metadata Generator - Git Repository. Available from: https://git.dcc.sib.swiss/sphn-semantic-framework/metadata-generator [accessed Nov 2, 2026]

16. Witte H, Unni DR, Krauss P, Touré V, Armida J, Österle S. The Swiss Personalized Health Network Metadata Catalog: Platform for Health Data Discovery and Exploration Based on Findable, Accessible, Interoperable, and Reusable Principles. JMIR Med Inform JMIR Publications, Toronto, Canada; 2026;14:e90146. doi: https://doi.org/10.2196/90146

17. Touré V, Krauss P, Gnodtke K, Buchhorn J, Unni D, Horki P, Raisaro JL, Kalt K, Teixeira D, Crameri K, Österle S. FAIRification of health-related data using semantic web technologies in the Swiss Personalized Health Network. Sci Data 2023;10(1):127. doi: 10.1038/s41597-023-02028-y

18. Touré V, Unni D, Krauss P, Marto AB, Kalt K, Stoira N, Pickl M, Österle S. SPHN Connector-A scalable pipeline for generating validated knowledge graphs from federated and semantically enriched health data. BMC Med Inform Decis Mak Springer; 2026;26(1):72.

19. OpenLink Virtuoso. 2025. Available from: https://virtuoso.openlinksw.com/knowledgegraph/ [accessed Nov 18, 2025]

20. Unni D, Witte H, Armida J, Touré V, Österle S. SPHN Metadata Catalog Schema. Available from: https://sphn.gitlab.io/sphn-metacat-schema/ [accessed Oct 30, 2026]

21. Krauss P, Touré V, Gnodtke K, Crameri K, Österle S. DCC Terminology Service—An Automated CI/CD Pipeline for Converting Clinical and Biomedical Terminologies in Graph Format for the Swiss Personalized Health Network. Applied Sciences 2021, Vol 11, Page 11311 Multidisciplinary Digital Publishing Institute; 2021 Nov 29;11(23):11311. doi: 10.3390/APP112311311

22. Bundesamt für Statistik (BFS). CHOP - Schweizerische Operationsklassifikation. 2025. Available from: https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/nomenklaturen/medk/instrumente-medizinische-kodierung.assetdetail.32128591.html [accessed Nov 18, 2025]

23. Regenstrief Institute Inc. UCUM - The Unified Code for Units of Measure. 2025. Available from: https://ucum.org/ucum [accessed Nov 18, 2025]

24. SPHN Metadata Catalog - FAIR Data Point. Available from: https://fdp.dcc.sib.swiss/ [accessed Jun 4, 2026]

25. SPHN FedData Analysis 2026.1 - Git Repository. Available from: https://git.dcc.sib.swiss/sphn-semantic-framework/outreach/sphn-feddata-analysis/-/tree/main/2026-1 [accessed Oct 30, 2026]

26. Touré V, Armida J, Unni D, Österle S. SPHN Federated Clinical Routine Dataset Metadata - Code Frequency and Concept Availability Oct 2026. Zenodo; 2026. doi: https://doi.org/10.5281/zenodo.23943915

27. Touré V, Armida J, Witte H, Unni D, Kalt K, Endrich O, Triep K, Gassner M, Kissling S, Vuaridel-Thurre G. Metadata-based Assessment of SNOMED CT and LOINC Adoption Across 800 000 Swiss Patients. 2026; doi: https://doi.org/10.21203/rs.3.rs-10377450/v1

687 28. GS1. GTIN - Global Trade Item Number. 2025. Available from:
688 https://www.gs1.org/standards/id-keys/gtin [accessed Nov 18, 2025]

689

690 List of abbreviations

691 ATC – Anatomical Therapeutic Chemical Classification System

692 CHOP – Swiss Classification of Hospital Operations

693 CT – Computed Tomography

694 DICOM – Digital Imaging and Communications in Medicine

695 ETL – Extract, Transform, Load

696 FAIR – Findable, Accessible, Interoperable, Reusable

697 GTIN – Global Trade Item Number

698 HGNC – HUGO Gene Nomenclature Committee

699 HUG – University Hospital Geneva

700 ICD-10-GM – International Statistical Classification of Diseases and Related Health
701 Problems, 10th Revision, German Modification

702 ICD-O-3 – International Classification of Diseases for Oncology, 3rd Edition

703 INSEL – University Hospital Bern

704 KISPI – University Children's Hospital Zurich

705 LOINC – Logical Observation Identifiers Names and Codes

706 MRI – Magnetic Resonance Imaging

707 ORDO – Orphanet Rare Disease Ontology

708 PET – Positron Emission Tomography

709 RDF – Resource Description Framework

SHACL – Shapes Constraint Language

SNOMED CT – Systematized Nomenclature of Medicine Clinical Terms

SPARQL – SPARQL Protocol and RDF Query Language

SPHN – Swiss Personalized Health Network

SPHN FedData – SPHN Federated Clinical Routine Dataset

UCUM – Unified Code for Units of Measure

USZ – University Hospital Zurich

Declarations

Ethics approval and consent to participate

Not applicable. Ethics approval was not required for this study because only metadata were analyzed, and no identifiable patient-level information was accessed or shared outside the hospital.

Consent for publication

Not applicable.

Availability of data and materials

The data used in this analysis, reflecting the state of the SPHN Metadata Catalog at the time of writing, have been deposited on Zenodo and are available at [26]. The R markdown script generated for building the different statistics and figures is available at [25].

The terminologies used in the analysis are hosted on the SPHN Terminology Service (https://portal.dcc.sib.swiss/terminologies). Access is restricted due to copyright and licensing constraints of some of the terminologies. The script used to transform selected terminologies into RDF format is publicly available at https://git.dcc.sib.swiss/dcc/biomedit-rdf.

Competing interests

The authors declare no competing interests.

Funding

This work was funded by the State Secretariat for Education, Research, and Innovation (SERI) via SPHN.

Authors' contributions

S.Ö and V.T. conceptualized the project. J.M led the data governance aspects related to the SPHN FedData, H.B, M.B, M.G, Y.J, M.J, K.K, J.V.I, H.P.T, G.V.T prepared the data in their respective hospitals, took care of the technical setup, maintained it and extracted the metadata. V.T and J.A performed the metadata analysis. S.O.E, J.A and V.T wrote the first draft of the manuscript. All authors have contributed to the manuscript and approved the final version.

Acknowledgements

The authors would like to thank all the teams in the hospitals, which contributed to making this data available in SPHN format. The authors would also like to thank all

colleagues involved, especially from USZ, Amanda Ramirez Ramos, Lirui Zhang and Daniele Vaccaro, at KISPI Beat Bangerter, at INSEL Pedram Bürgin.

Tables

Table 1. SPHN concepts available in hospitals grouped into specific domains. A total of 68 core concepts and 73 supporting concepts defined in the SPHN RDF Schema have been grouped into higher-level categories to facilitate readability in the analysis' plots.

Domains	Core concepts	Supporting concepts
Administrative	Administrative Case, Consent, Follow Up, Healthcare Encounter, Insurance Status, Resuscitation Directive	Admission; Discharge; Care Handling; Location; Therapeutic Area; Vital Status
Clinical condition	Access Device Presence, Allergy, Assessment Event, Billed Diagnosis, Diagnosis, Implant Presence, Nursing Diagnosis, Problem Condition, Tobacco Exposure, Transplant Presence	Access Device; Implant; Transplant; Allergen; Assessment; Assessment Component; Assessment Result
Demographics	Administrative Sex, Age, Birth, Civil Status, Death, Home Address, Nationality	Birth Date; Death Date; Country; Swiss Socio Economic Position; Subject Pseudo Identifier
Imaging	Computed Tomography Imaging Series, Contrast Agent Administration Event, Imaging Series, Imaging Procedure, Magnetic Resonance Imaging Series, Positron Emission Tomography Imaging Series, Radiopharmaceutical	Imaging Frame; Computed Tomography Imaging Frame; Magnetic Resonance Imaging Frame; Positron Emission Tomography Imaging Frame; X Ray Imaging Frame; Image Dimensions

	Administration Event, X Ray Imaging Series	
Laboratory	Antimicrobial Susceptibility Lab Test Event, Biobanksample, Isolate, Lab Test Event, Microbiology Biomolecule Presence Lab Test Event, Microbiology Microscopy Lab Test Event, Microorganism Identification Lab Test Event, Sample, Sample Processing	Lab Test; Lab Result; Antimicrobial Susceptibility Lab Test; Antimicrobial Susceptibility Result; Microbiology Biomolecule Presence Lab Test; Microbiology Biomolecule Presence Result; Microbiology Microscopy Lab Test; Microbiology Microscopy Result; Microorganism Identification Lab Test; Microorganism Identification Result; Organism; Reference Range; Reference Value; Gene
Measurement	Blood Pressure Measurement, Body Height Measurement, Body Mass Index, Body Position, Body Surface Area, Body Temperature Measurement, Body Weight Measurement, Cardiac Index, Cardiac Output Measurement, Circumference Measurement, Fluid Balance, Fluid Input Output, Gestational Age At Birth, Heart Rate Measurement, Nutrition Intake Event, Oxygen Saturation Measurement, Respiratory Rate Measurement, Time Series Data File	Blood Pressure; Body Height; Body Temperature; Body Weight; Cardiac Output; Circumference; Heart Rate; Oxygen Saturation; Respiratory Rate; Physiologic State; Nutritional Product
Medical procedure	Billed Procedure, Electrocardiographic Procedure, Medical Procedure, Organ Support, Radiotherapy Procedure	
Medication and treatment	Drug Administration Event, Drug Prescription Event, Oxygen Administration Event	Drug Prescription; Drug; Drug Article; Pharmaceutical Dose Form; Substance; Time Pattern; Fraction Of Inspired Oxygen

Oncology	Oncology Diagnosis, Oncology Surgery, Tumor Grade Assessment Event, Tumor Stage Assessment Event	Tumor Grade Assessment; Tumor Grade Assessment Result; Tumor Stage Assessment; Tumor Stage Assessment Result
Provenance		Source System; Healthcare Primary Information System; Data Provider; Department; Data Release; Data Determination
Cross-domain concepts		Code; Quantity; Unit; Body Site; Laterality; Medical Device; Intent

765

766 **Table 3. Evolution of key statistics of the SPHN FedData between versions 2025.1 and**
767 **2026.1.** The table summarizes the changes in the number of participating hospitals, SPHN
768 concepts, core concepts, shared concepts across hospitals, triples, distinct terminology codes,
769 and instances. Newly delivered and removed concepts are indicated where applicable in the
770 “delta” column. Concepts with an asterisk (*) indicate that their presence in the “delta” column
771 is due to concept remodeling in 2026.1. Bn = billion.

Metric	2025.1	2026.1	Delta
Participating hospital	6	5	- University Hospital Basel
Patient count	815084	959978	+17.8%
SPHN concepts (all)	125	141	+ Computed Tomography Imaging Frame, Computed Tomography Imaging Series, Drug Prescription Event, Image Dimensions, Imaging Frame, Imaging Series, Magnetic Resonance Imaging Frame, Magnetic Resonance Imaging Series, Medical Procedure, Nutrition Intake Event, Nutritional Product, Positron Emission Tomography Imaging Frame, Positron Emission Tomography Imaging Series, Problem Condition, Tobacco Exposure, X Ray Imaging Frame, X Ray Imaging Series - Nutrition Intake (replaced by Nutrition Intake Event)

SPHN core concepts	60	68	+ Computed Tomography Imaging Series, Drug Prescription Event*, Imaging Series, Magnetic Resonance Imaging Series, Medical Procedure, Nutrition Intake Event, Positron Emission Tomography Imaging Series, Problem Condition, Tobacco Exposure, X Ray Imaging Series - Drug Prescription*, Nutrition Intake
Core concepts shared by all hospitals	18	26	+ Allergy, Blood Pressure Measurement, Body Mass Index, Body Weight Measurement, Circumference Measurement, Electrocardiographic Procedure, Gestational Age At Birth, Microorganism Identification Lab Test Event, Nationality, Respiratory Rate Measurement. - Drug Prescription, Oxygen Saturation Measurement.
Triples [x10^9]	12.1	16.5	+ 36.6 %
Distinct codes	41140	64800	+ 57.5 %
Terminology instances	750 M	803 M	+ 7.2 %

772