Data Harmonisation for Gynaecological Cancer Research

- R. G. Hamed, O. Ibrahim, L. McMahon, S. Hurley, F. Mangan, J. J O'Leary, A. Long, S. O'Toole, C. M Martin on behalf of SEARCH Consortium; Trinity St James Cancer Institute (TSJCI), Trinity Translational Medicine Institute ,Trinity College Dublin (TCD), Ireland

FACTS & FIGURES

Start date:	01 Oct 2024
End date:	30 Sep 2028
IHI funding:	7,916,214 €
Industry contribution:	2,753, 005 €
Contributing partners:	3,733,843 €
Total Cost:	14,403,062 €
Project website:	ihi-search.eu
Social media:	@IHISEARCH

CHALLENGE

From synthetic data to harmonisation

- SEARCH aims to unlock healthcare data for research and innovation by generating privacy-preserving synthetic datasets through a federated learning approach, without centralising or exposing real world patient-level data.
- The utility of synthetic data depends directly on the quality, structure and interoperability of the real source data used to train the generators.
- Gynaecological cancer data are especially heterogeneous; cervical screening carries overlapping classifications and HPV coding; ovarian datasets mix clinical, genomic and treatment variables.
- Without harmonisation, poorly standardised inputs propagate noise and bias into synthetic outputs and downstream federated AI models.

APPROACH & METHODOLOGY

1 Initiation
Leveraged two rich research and clinical datasets: CERVIVA cervical screening data (>12,000 women; biomarkers and long-term follow-up) and TSJCI Biobank ovarian cancer data (structured and genomic data).

2 Requirement analysis
Profiled source datasets and data dictionaries; assessed source terminologies and structures; validated key variables with clinical experts; identified information-loss and unmapped-concept risks.

3 Transformation design
Defined OHDSI-aligned transformation rules, mapped source concepts using Athena and Usagi, and aligned source data to appropriate OMOP CDM tables and fields.

4 Implementation & QA
Performed cleaning, deduplication, missing-data handling and de-identification; executed Python ETL; validated mappings and cross-field consistency to prepare harmonised data for federated model training.

5 Maintenance
Established periodic review and update plans for vocabulary changes, mapping refinements and evolving source data.

CAREER IMPACT

Building on a PhD in computer science and a data science background, SEARCH has broadened my expertise in three complementary areas:

- **Clinical data & oncology context**
- **Interoperability & standards**
- **Governance, ethics & responsible AI**

RESULTS

Results at a glance

14,213 patients
12,768 cervical + 1,445 ovarian

98 clinical fields
50 CERVIVA + 49 TSJCI

200+ source values mapped
to standard clinical terminologies

4 Python QA scripts
cleaning, consistency, validation & integration

Additional genomic depth: whole-genome sequencing (WGS) will be available for 50 ovarian cancer samples.

Harmonisation & data-quality findings

- **Missing data**: Sentinel nulls harmonised; expected structural nulls distinguished from true gaps.
- **Cross-field inconsistencies**: Treatment, BRCA, survival (PFS/OS) and laboratory relationships validated in the ovarian dataset.
- **Data standardisation**: CINtecPLUS in cervical dataset and free-text fields in ovarian dataset converted to structured categories in the cervical dataset.
- **Data-entry errors**: Spelling variants, HPV platforms and colposcopy codes corrected and normalized in the cervical dataset.

Standardisation coverage

SNOMED CT · LOINC · ICD-O-3 · RxNorm · Bethesda · FIGO

WHAT THIS ENABLES FOR SEARCH

Harmonisation provides the foundation for subsequent synthetic-data generation and research

Heterogeneous cancer data → Harmonised OMOP data → Federated synthetic data generation & QA → Research-ready synthetic data

The harmonisation work creates a reproducible, standards-based foundation for downstream synthetic-data generation, validation and AI use across SEARCH gynaecological cancer studies.

VALUE OF IHI COLLABORATION

The IHI partnership combines complementary clinical, technical and industry expertise, with selected partners contributing to this work within the wider consortium.

Clinical & Data Expertise	**TCD (lead):** study design, data stewardship and OMOP/ETL
Methods & Platform Expertise	**UTH:** synthetic data generation **Maggioli:** federated platform
Industry & Implementation	**IQVIA, IBM, Medicalvalues:** OMOP/terminology harmonisation **BYTE:** infrastructure

TAKE-HOME MESSAGE

Harmonisation turns fragmented cancer data into reusable, interoperable evidence.

By standardising structure, terminology and quality, SEARCH creates a consistent foundation for cross-study analysis, synthetic data generation, validation and responsible AI research.

This project is supported by the Innovative Health Initiative Joint Undertaking and its members.