De-identification and Quality Control steps to comply with data authorities and create a clinical grade CDM

**INTRO:**
- Standardizing registry data to address multifactorial diseases like colorectal cancer.
- Enriching the standardization pipeline with additional layers of patient de-identification and quality control in order to comply with data authorities and produce a clinical grade CDM that can be used for direct patient interventions.

**METHODS**

**Data:** DCCG national clinical registry initiated at 2001, with 384 variables, ~77k patients.

**Transformation process:**
- **Data Curation:**
  1. invalid dates, 2. illogical values
- **De-identification:**
  1. hash direct ids, 2. remove infrequent events, 3. person date shift, 4. combine infrequent ages, 5. modify physiological measurements.
- **OMOP ETL:**

**Evaluation of transformation:**
- **DC&De-id:**
  1. column-based logging with number of affected records from each rule, 2. statistical comparison and medical evaluation.
- **OMOP ETL:**
  1. statistical comparison, 2. column-based logging, 3. random patient spot-check, 4. Data Quality Dashboard.
- **Overall assessment:** descriptive statistics of 16 selected variables between source, de-identified and CDM data and association with 30-day post-operative death using logistic regression.

**RESULTS**

- Mapping coverage of 99.12% (85% standard and 15% custom mappings).
- 76,849 rows and 317 variables to 9,668,672 records in OMOP CDM without losing any patients.
- 22,869 data points were lost because of missing dates.
- DC&De-id steps did not affect the general distributions of the data.
- Overall assessment showed no significant difference between source, de-id and CDM data.