Federated learning and federated queries using head and neck cancers as a use case

Laura Botta

Evaluative Epidemiology Unit, Department of Epidemiology and Data Science, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy

laura.botta@istitutotumori.mi.it

L'ONCOLOGIA ITALIANA È NATA QUI

Fondazione IRCCS Istituto Nazionale dei Tumori
via Venezian, 1 20133 Milano

Sistema Socio Sanitario
Regione Lombardia

Where are we?

Individual pseudonymised data
Population-Based Cancer Registry (PBCR) Experience

EUROCARE study
Survival of cancer patients in Europe

About 100 PBCRs involved, 30 countries.
In the future: data sharing at individual level is uncertain, possible solutions will be privacy assessments between the PBCR and JRC and INT/ISS or hybrid analysis (using pooled data).

BENCHISTA project
International benchmarking of childhood cancer survival by stage

About 70 PBCRs involved. To achieve research collaboration: **18** months to finalize the privacy assessment.
In future: all this work will have to be redone.

References:
Long-term survival and cure fraction estimates for childhood cancer in Europe (EUROCARE-6): results from a population-based study. Botta et al. LO 2022
International benchmarking of childhood cancer survival by stage at diagnosis: The BENCHISTA project protocol. Botta et al. PLOS ONE 2022
Cancer data quality and harmonization in Europe: the experience of the BENCHISTA Project. Lopez-Cortes et al. Frontiers 2023

Solutions?

Federated queries and Learning

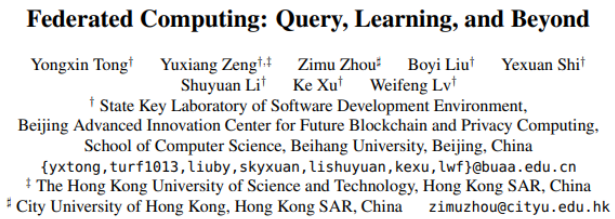
Federated Query

Purpose: Federated queries are used to retrieve and integrate data from multiple, distributed sources as if they were a single database.
Operation: When a federated query is executed, it sends sub-queries to different data sources, collects the results, and combines them into a unified response. This is particularly useful for data integration and analysis across various databases without moving the data.

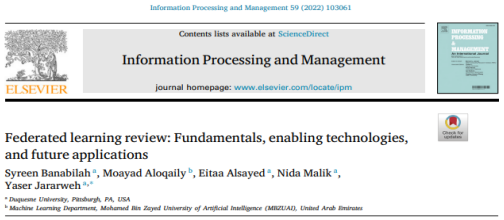
Federated Learning

Purpose: Federated learning is a machine learning approach where a model is trained across multiple decentralized devices or servers holding local data samples, without exchanging the data itself.
Operation: Each device trains the model locally on its data and only shares the model updates (like gradients or weights) with a central server. The server aggregates these updates to improve the global model, which is then redistributed to the devices for further training.

- Autonomous Constraint: any data owner does not share his raw data to others.
- Security Constraint: during the computation, except for the result, any sensitive data of a data owner cannot be leaked to others.


Federated Computing: Query, Learning, and Beyond

Yongxin Tong† Yuxiang Zeng†,‡ Zimu Zhou♯ Boyi Liu† Yexuan Shi†
Shuyuan Li† Ke Xu† Weifeng Lv†
† State Key Laboratory of Software Development Environment,
Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing,
School of Computer Science, Beihang University, Beijing, China
{yxtong,turf1013,liuby,skyxuan,lishuyuan,kexu,lwf}@buaa.edu.cn
‡ The Hong Kong University of Science and Technology, Hong Kong SAR, China
♯ City University of Hong Kong, Hong Kong SAR, China zimuzhou@cityu.edu.hk


Information Processing and Management 59 (2022) 103061
Contents lists available at ScienceDirect
Information Processing and Management
journal homepage: www.elsevier.com/locate/ipm

Federated learning review: Fundamentals, enabling technologies, and future applications

Syreen Banabilah [a], Moayad Aloqaily [b], Eitaa Alsayed [a], Nida Malik [a], Yaser Jararweh [a,*]
[a] Duquesne University, Pittsburgh, PA, USA
[b] Machine Learning Department, Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), United Arab Emirates

Example of federated queries algorithm

(Centralized) Average

Age
57
68
32
47
28

$$\mu = \frac{232}{5} = 46.4$$

Federated Average

Age	
57	Organization A
68	
32	
47	Organization B
28	

$$\mu = \frac{(157 + 75)}{3 + 2} = 46.4$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i$$

$$\mu = \frac{1}{n_a + n_b}\left(\sum_{i=1}^{n_a} \vec{x}_{a,i} + \sum_{i=1}^{n_b} \vec{x}_{b,i}\right)$$

Federated queries and learning approach

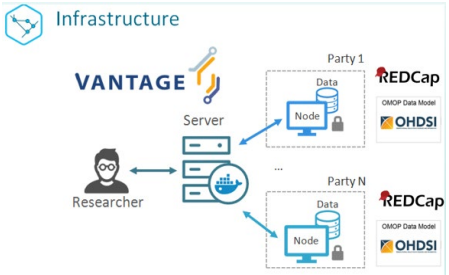
Population based cancer registry data:
RARECARENET Asia

PBCR Head and neck data analyzed using VANTEGE6 An open-source infrastructure for privacy preserving analysis.

Clinical Cancer Registry:
STARTER

STARTER PROJECT
STarting an Adult Rare Tumour European Registry

Hospital based cancer registry collecting rare head and neck cancer data. VANTEGE6. Legal framework and DPIA in place that will last “forever”.



References:
Head and neck cancers survival in Europe, Taiwan, and Japan: results from RARECAREnet Asia based on a privacy-preserving federated infrastructure. Botta et al. Frontiers Oncology 2023
The observational clinical registry of the ERN on Rare Adult Solid Cancers: The protocol for the rare head and neck cancers. Trama et al. PLOS ONE 2022

Which algorithms are feasibile right now in the federated setting of the H&N registry:

- **Descriptive analyses of continuous and categorical variables**
    - for continuos: mean, median and interquartile range,
    - for categorical: frequency distribution, **Contingency table (**with row and column total and percentage)
- **Chi-squared test χ^2** for categorical variables
- **T-test** for continuous variables
- **Generalized linear models, GLM** (e.g., Logistic regression with continuous or categorical covariates)
- **Kaplan-Meier survival analysis** stratified by a categorical variable.
- **Log Rank test** to assess the differences in Kaplan-Meier survival strata
- **Cox multivariable model** adjusted by categorical variables (e.g., prognostic factors).
- **Schoenfield residuals** to test the Cox model proportional hazard assumption of the Cox multivariable model

Data quality

As you will never see the row data you have to be confident that the quality of each DB included in the analysis is checked and corrected, if needed.

How we are dealing with this checks in the federated registry?

1. The coordinating centre decides to check the quality of the data and runs an R script that executes it locally.

2. The data quality results are sent to the coordinator for evaluation.

3. A report is provided to all participating centres including the completeness of the variables, the list of cases with errors and some preliminary tables to check consistency.

4. Errors are checked and resolved by the data owner.

5. The coordinator verify this revision running again the same data quality check

To ensure standardization:

Training of the data managers
Recording a training session on how to use the CRF and example of fictictious cases to be included in the CRF by our clinical expert
Writing a codebook and making it available on the web site.

Features	Centralised vs federated
Latency of computation	CENTRALIZED reduced reliance on external systems; in the federated analysis the speed is set to the slower machine involved
Data management/Data analysis	CENTRALIZED Individual level data quality checks; all type of analysis are feasible; Possibility to aggregate countries to overcome rarity issue
Lightness of technical implementation	CENTRALIZED IT infrastructure needed is easier
Data updated	BOTH
Data availability	BOTH FEDERATED data are always accessible when needed but the CENTRALIZED relies less on external sources
Privacy assessment	FEDERATED is more privacy preserving
Security / Data breach	FEDERATED Reduced amount of data in case of breach
Privacy-by-design principles	FEDERATED Avoids creating additional copies of data, stored in the original source system and does not have to be communicated or transfer
Expanding trust	FEDERATED Possibility to opt in/out; all analyses and requests are tracked.

Where are we going?

- Regardless the type of data (population based /clinical data): Privacy assessment is part of research life and we can’t ignore it

- Peculiarities of rare cancers

Although the IT infrastructure required is complex, the FEDERATED LEARNING APPROACH is evolving rapidly.
It is THE FUTURE, but it takes time.

PBCRs could be in my opinion the first to benefit from the federated approach because they are dedicated to research (technical readiness, standardized data collections).

IN THE MEANTIME:

- Standardized the dataset as much as possible across countries and projects

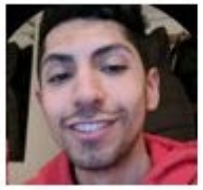
- Using an Hybrid model (individual data + grouped data) if possible. Ok for some statistical analysis (descriptive analysis, univariate models) but difficult for others such as multivariate analysis and Propensity score definition. Difficult for data research/exploration.

Many thanks to you and to all the people who collaborated with me on these projects














CHILDREN WITH CANCER UK

AIRC

European Commission

laura.botta@istitutotumori.mi.it