Fine-tuning OMOP databases to create Analysis-Ready Data is a key step in ensuring robust insights for small cohorts in a federated setting

Title: Federated Analytics on OMOP-CDM data: evaluating dsOMOP and complementary methods to optimize reproducibility of descriptive statistics on small cohorts

Background: Interoperability between healthcare databases is increasingly important for conducting large-scale observational studies from multiple sources. While the OMOP Common Data Model (CDM) is widely used to standardize data structure and semantics, hospitals often remain reluctant to centralize patient-level data due to strict privacy regulations. Federated analytics offers a solution by bringing the analysis directly to the local data. Recently, the dsOMOP package (1) was introduced to bridge this gap, connecting OMOP-formatted databases with DataSHIELD, a secure infrastructure for non-disclosive federated analysis.

Methods

The small KADOr dataset (N=315) was standardized to OMOP (2) and distributed across 3 PostgreSQL nodes (N=157, N=94, N=64) connected to a DataSHIELD server (Figure 1). This small dataset represents a challenging scenario for standard privacy controls.

To reproduce 72 centralized descriptive statistics and subgroups analysis (by pCR and by treatments), we compared 3 data access workflows:

- **Standard dsOMOP**: Queries via the translation layer with extraction-level privacy checks.
- **dsOMOP + Aggregated Concepts:** Injection of grouped concepts upstream to increase patient density.
- **Direct Access:** Bypassing the translation layer using the resourcer package (3) to apply disclosure controls at the aggregation step only.

Performance was assessed using the Exact Match Rate (EMR) against centralized results across varying privacy thresholds (nfilter in {1,2,3} defined as the minimum non-zero count of observational units, typically individuals, in a subset).

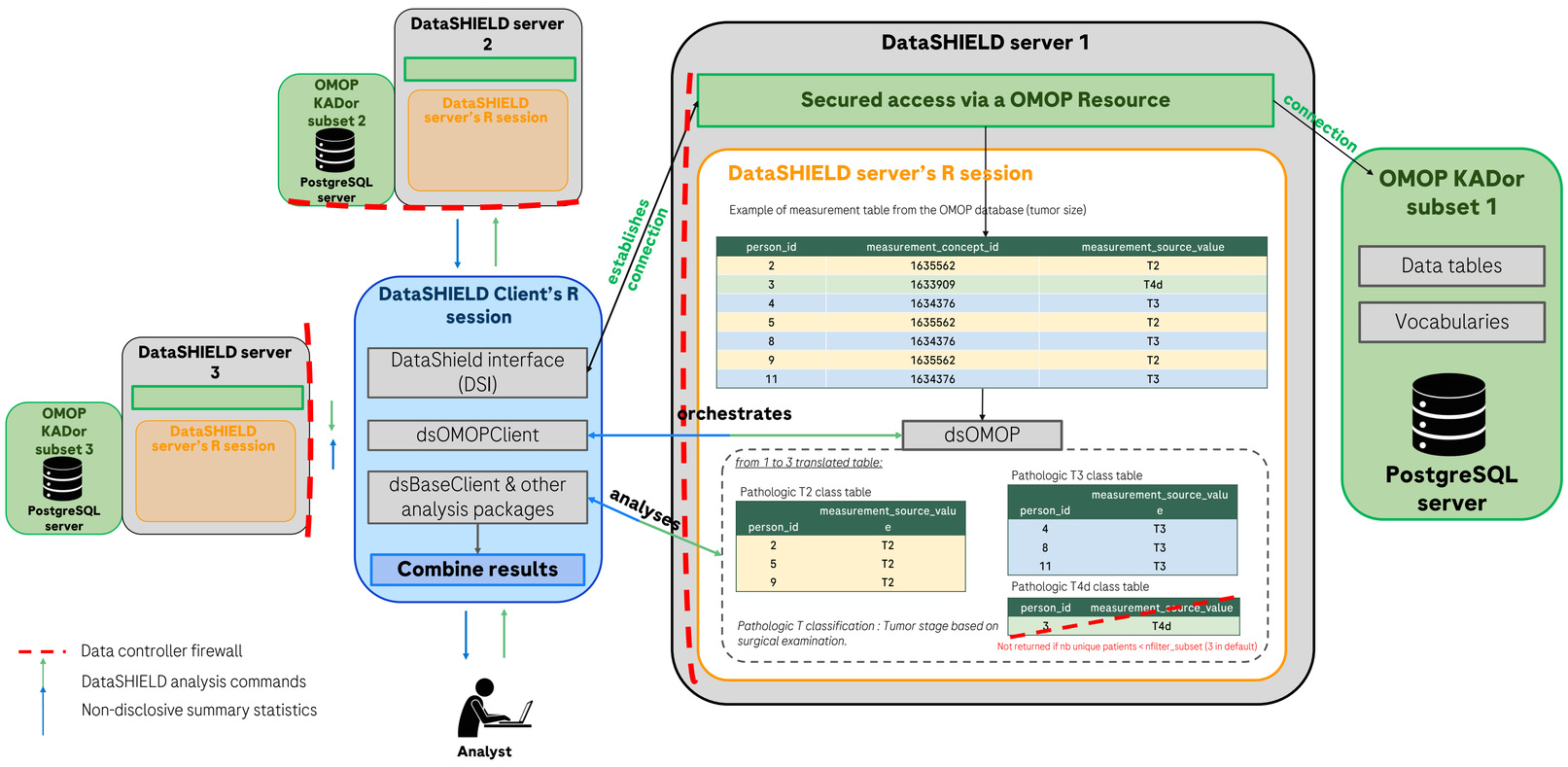

Figure 1. Federated infrastructure data flow

	dsOMOP	+ new concepts	+ new resource	
nfilter_subset & tab = 3	44%	66%	66%	With subgroups analysis
nfilter_subset & tab = 2	46%	69%	69%	
nfilter_subset & tab = 1	100%	100%	100%	
nfilter_subset & tab = 3	67%	67%	71%	Without subgroups analysis
nfilter_subset & tab = 2	72%	72%	72%	
nfilter_subset & tab = 1	100%	100%	100%	

Table 1. Exact Match Rates by method and confidentiality threshold

Results

Under default privacy settings (nfilter=3), the standard dsOMOP approach demonstrated limited reproducibility, achieving an EMR of only 44% (Table 1). The primary bottleneck was the extraction-level (using nfilter_subset & tab), which systematically blocked granular concepts required for subgroup definitions such as pathologic complete response (pCR) (Figure 1 - pCR T4d class example).

To overcome this, we injected pre-aggregated concepts upstream directly into the measurement table, which improved the EMR to 66%. However, neither this data engineering approach nor the use of Direct Access (via the resourcer package to bypass extraction-level filters) was sufficient to achieve full concordance under default settings. In our case, a perfect 100% EMR—faithfully reproducing all 72 analyses—could only be achieved by lowering the privacy filters to nfilter=1.

Limitations: This highlights a critical methodological limitation: when working with small datasets researchers must carefully anticipate their analytical strategy. DataSHIELD's privacy disclosure traps inherently restrict the feasibility of highly granular subgroup analyses on small cohorts, demonstrating that not all desired analyses can be technically executed without relaxing standard data protection | privacy settings.

Conclusion: While standard dsOMOP guarantees robust privacy protection, its default filters restrict analytical depth in small cohorts. We strongly suggest to researchers to careful monitor data and privacy levels required to answer research objectives. A feasibility study should be performed:
1/ to evaluate if additional concepts are necessary (what we call ARD: Analysis Ready Data), aligned with each scientific assessment criteria.
2/ to estimate statistical precision obtained according to different data filters level in each node (or center).

(1) Sarrat-González D, Escribà-Montagut X, Houghtaling J, González JR. dsOMOP: bridging OMOP CDM and DataSHIELD for secure federated analysis of standardized clinical data. Bioinformatics. 2025 Jun 2;41(6):btaf286. doi: 10.1093/bioinformatics/btaf286. PMID: 40327502; PMCID: PMC12187060.
(2)Lambert A, Castagne C, Pau D, Chmiel J, Labarga A, Boernert E, et al. Evaluating the statistical utility and information loss in the transformation of a real-world oncology database from CDISC-SDTM to OMOP-CDM. ESMO Real World Data Digit Oncol. 2026 Mar1;11:100655.
(3) Yannick Marcon, Tom Bishop, Demetris Avraam, Xavier Escribà-Montagut, Patricia Ryser-Welch, Stuart Wheater, Paul Burton, Juan R González. Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. *Plos Comp Biol*. 2020

Tristan Laurent[1], Claire Castagne[1], Jacek Chmiel[5], Eric Boernert[2], François Margraff[2], Camille Bachot[1], Lukasz Kaczmarek[2], Dimitar Toshev[2], Judith Martinez Gonzalez[7], Thomas Stone[4], David Pau[1]

[1]Roche SAS, Boulogne-Billancourt, France. [2]F. Hoffmann-La Roche AG, Basel, Switzerland. [3]Roche Polska Sp. z o.o., Warsaw, Poland. [4]Roche Products Limited, Welwyn Garden City, United Kingdom. [5]Avenga, Cologne, Germany. [7]Billennium S.A., Warsaw, Poland.