

Virtual Pooling: Accurate, Scalable, and Privacy-Preserving Analytics Without Aggregating Patient Data

Kunlong Liu^{1,2}, Goktug Onal⁴, Ishtiyaque Ahmad^{1,3}, Vivek Rudrapatna⁴, Trinabh Gupta^{1,2}

1 DataUnite, Inc.

2 UC Santa Barbara

3 UC Santa Cruz

4 UC San Francisco

PRESENTED BY

Ishtiyaque Ahmad
CTO, DataUnite, Inc.

Disclosure of financial relationship: Ishtiyaque Ahmad is an employee of and holds equity in DataUnite, Inc., the developer of the Virtual Pooling technology presented here.

Two major approaches for multicenter studies

Direct Pooling

- ✓ **Gold-standard accuracy**
- ✗ High governance burden
- ✗ Risk to patient privacy

Distributed Research Networks

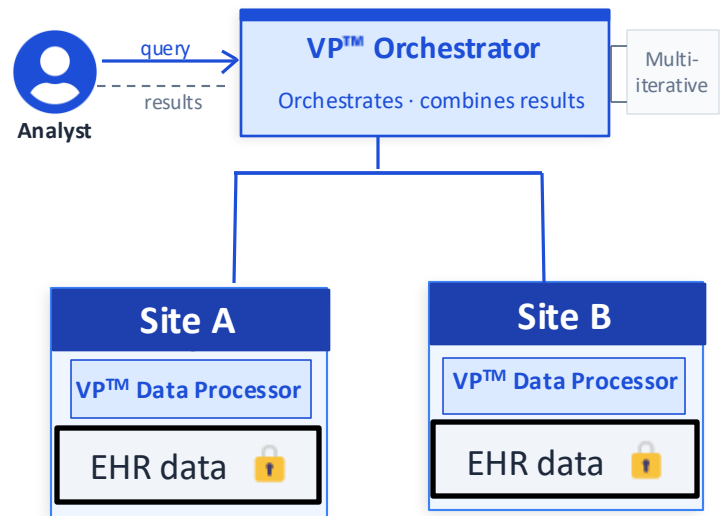
Federated networks: OHDSI, PCORnet, Sentinel, PDA, etc.

- ✓ **Preserves patient privacy**
- ✗ Accuracy compromised by meta-analysis / distributed algorithms¹
- ✗ Not robust to skewed data distributions²

How can we achieve the best of both worlds?

1. Lu et al. *Meta-Analysis and Federated Learning over Decentralized Distributed Research Networks*. Annual Review of Biomedical Data Science, 2025.
2. Higgins et al. *Quantifying heterogeneity in a meta-analysis*. Statistics in medicine, 2002

Virtual Pooling: Multicenter analytics with *privacy* and *accuracy*



Key idea

A federated platform

- ✓ Ensures patient privacy
- ✓ Lowers the governance barrier

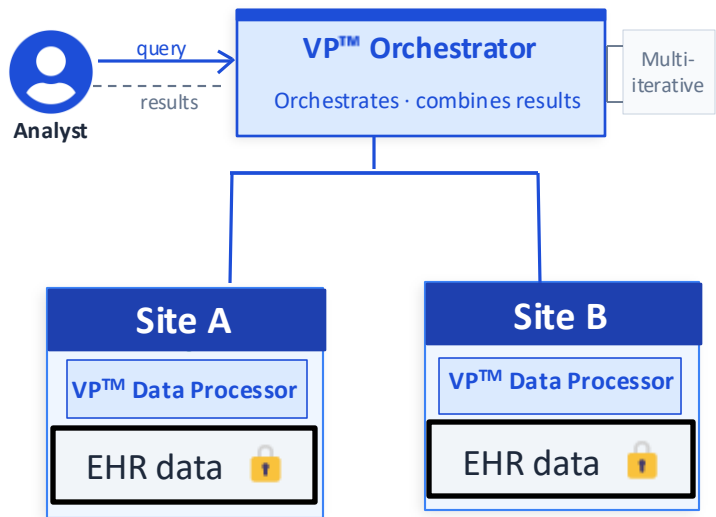
Specialized federated algorithms for biostatistical operations

- ✓ Accuracy equivalent to pooled analysis
- ✓ Robustness to distribution skew

Virtual Pooling: Multicenter analytics with *privacy* and *accuracy*

1 Distributed Infrastructure

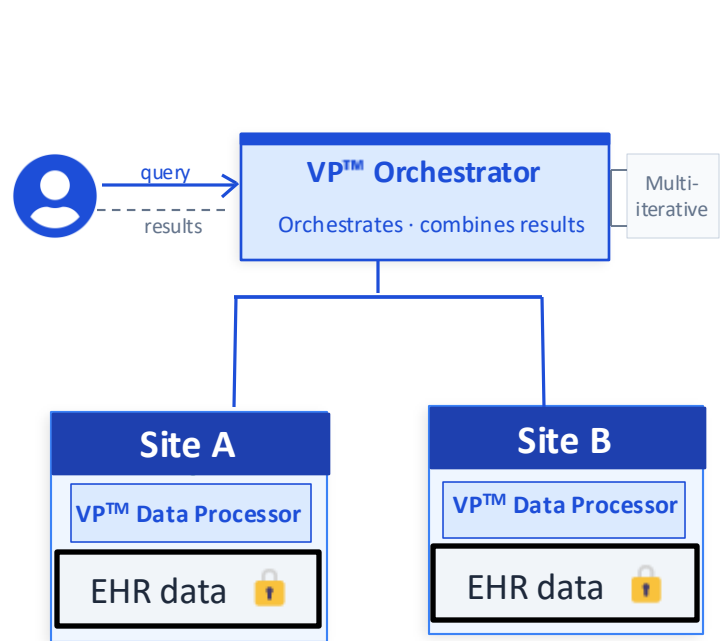
2 Data Harmonization Layer



3 Federated Analytics Algorithms

4 Interactive User Interface

Virtual Pooling: Multicenter analytics with *privacy* and *accuracy*



1 Distributed Infrastructure

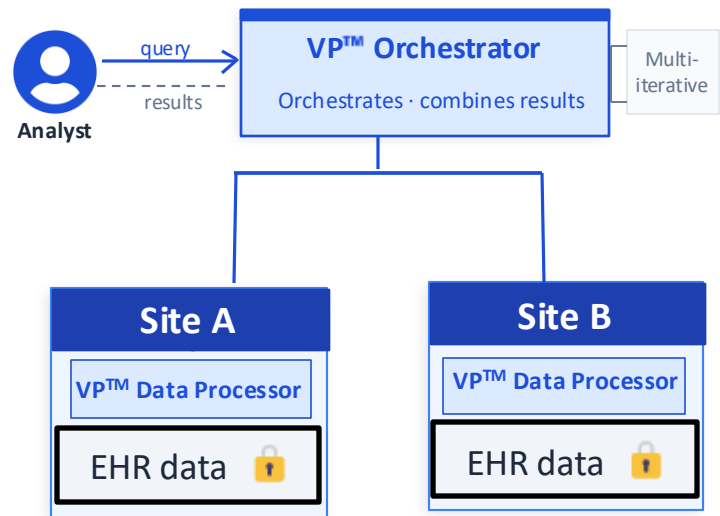
- **VP Orchestrator:** hosted in the cloud
- **VP Data Processor:** runs at each site
- No patient level data leaves site
- Iterative multi-round operations
- Secure and privacy-compliant

2 Data Harmonization Layer

3 Federated Analytics Algorithms

4 Interactive User Interface

Virtual Pooling: Multicenter analytics with *privacy* and *accuracy*



1 Distributed Infrastructure

- **VP Orchestrator:** hosted in the cloud
- **VP Data Processor:** runs at each site
- No patient level data leaves site
- Iterative multi-round operations
- Secure and privacy-compliant

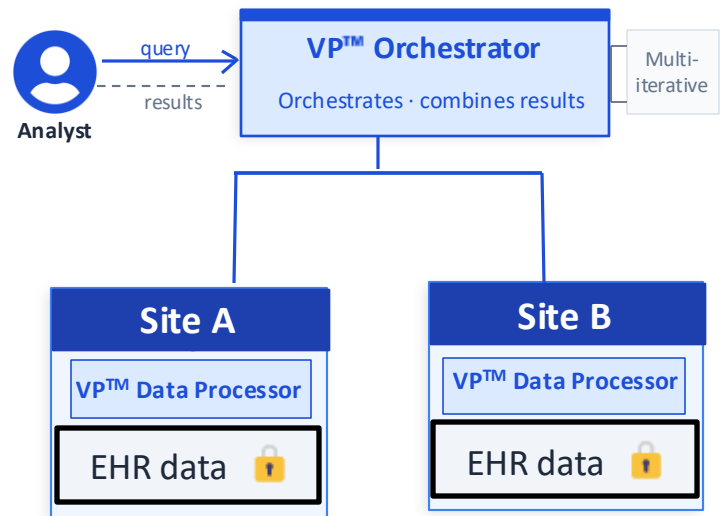
2 Data Harmonization Layer

- Schema & column mapping
- Type normalization
- Uniform category encoding
- Missing value imputation
- And more...

3 Federated Analytics Algorithms

4 Interactive User Interface

Virtual Pooling: Multicenter analytics with *privacy* and *accuracy*



1 Distributed Infrastructure

- **VP Orchestrator:** hosted in the cloud
- **VP Data Processor:** runs at each site
- No patient level data leaves site
- Iterative multi-round operations
- Secure and privacy-compliant

2 Data Harmonization Layer

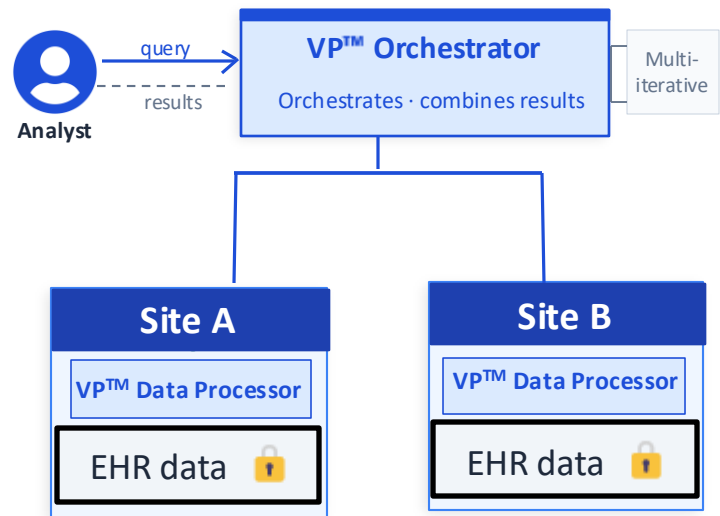
- Schema & column mapping
- Type normalization
- Uniform category encoding
- Missing value imputation
- And more...

3 Federated Analytics Algorithms

- **Results identical to pooled analysis**
- Descriptive stats, hypothesis testing
- Generalized linear models
- Propensity score matching
- Kaplan-Meier curve, Cox regression
- Causal inference (ATE, ATT), etc.

4 Interactive User Interface

Virtual Pooling: Multicenter analytics with *privacy* and *accuracy*



1 Distributed Infrastructure

- **VP Orchestrator:** hosted in the cloud
- **VP Data Processor:** runs at each site
- No patient level data leaves site
- Iterative multi-round operations
- Secure and privacy-compliant

2 Data Harmonization Layer

- Schema & column mapping
- Type normalization
- Uniform category encoding
- Missing value imputation
- And more...

3 Federated Analytics Algorithms

- **Results identical to pooled analysis**
- Descriptive stats, hypothesis testing
- Generalized linear models
- Propensity score matching
- Cox survival analysis
- Causal inference (ATE, ATT), etc.

4 Interactive User Interface

- Single analytical interface
- Feels like working with a local dataset
- Zero infrastructure management
- Interactive and incremental analysis
- No upfront pipeline required

We validated Virtual Pooling (VP) by replicating a published N3C COVID-19 study* consisting data from 59 sites

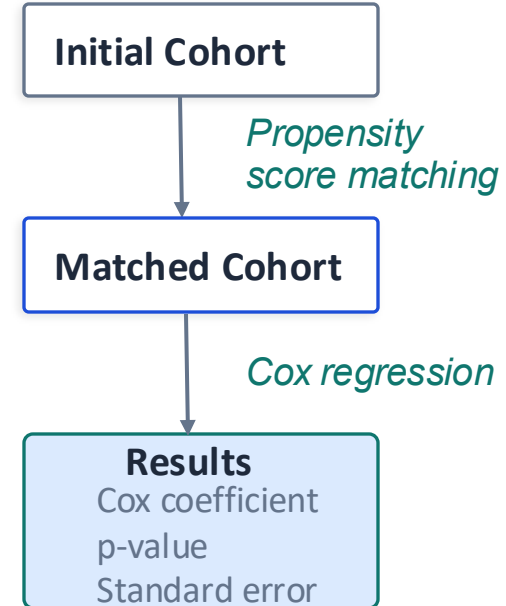
The study

- Association between chronic immunosuppressant use and in-hospital outcomes in COVID-19 patients
- Data: National COVID Cohort Collaborative (N3C) — secure, NIH-managed, HIPAA-limited enclave
- 59 sites: initial cohort of 334,754 patients

Validation questions

1. How accurate is VP compared to the gold standard of direct pooling?
2. How robust is VP under skewed data distributions across sites?

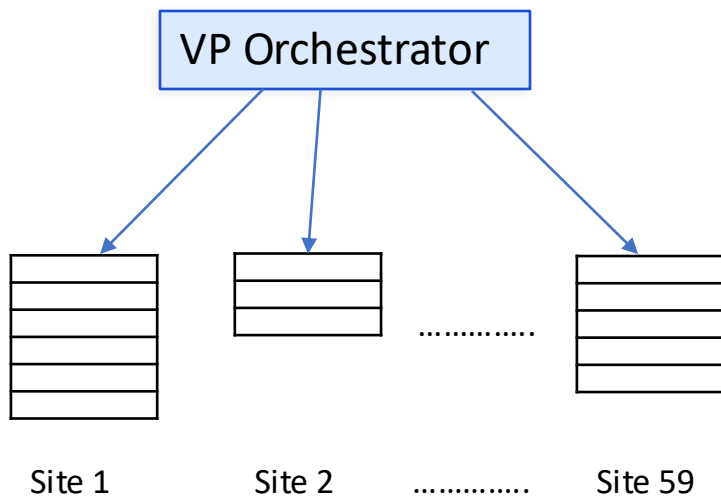
Analytical workflow



* Andersen et al. *Long-term use of immunosuppressive medicines and in-hospital COVID-19 outcomes: a retrospective cohort study using data from the National COVID Cohort Collaborative*. The Lancet Rheumatology 4, no. 1 (2022): e33-e41.

How accurate is VP compared to the gold standard of direct pooling?

Step 1: Federated propensity score matching

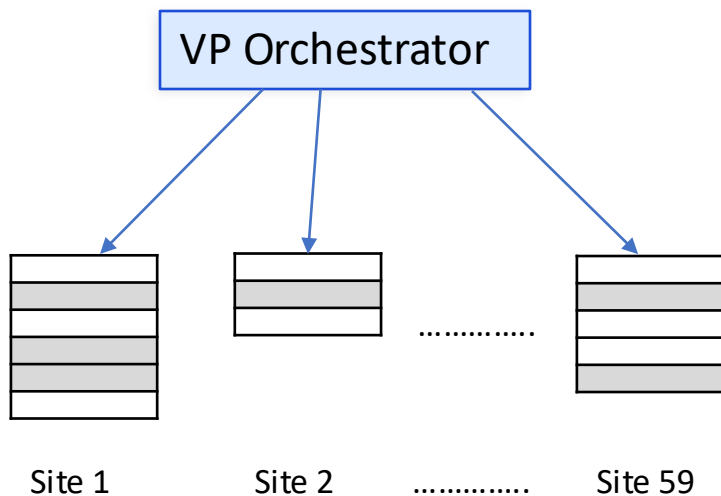


Procedure

- Federated logistic regression to estimate propensity scores
- Nearest-neighbor matching
(4-to-1 NN, 0.2 SD caliper)

How accurate is VP compared to the gold standard of direct pooling?

Step 1: Federated propensity score matching



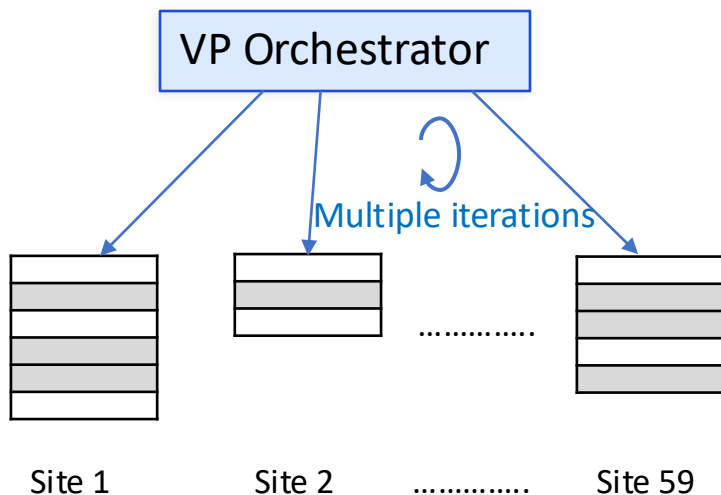
Procedure

- Federated logistic regression to estimate propensity scores
- Nearest-neighbor matching
(4-to-1 NN, 0.2 SD caliper)

VP identified the *exact same matched cohort* of 80,767 patients as the direct pooled analysis.

How accurate is VP compared to the gold standard of direct pooling?

Step 2: Federated Cox proportional hazards model (51 covariates)



Procedure

1. Compute local gradients and Hessians
2. Cryptographically mask local statistics
3. Securely aggregate across sites
4. Update coefficients and iterate to convergence

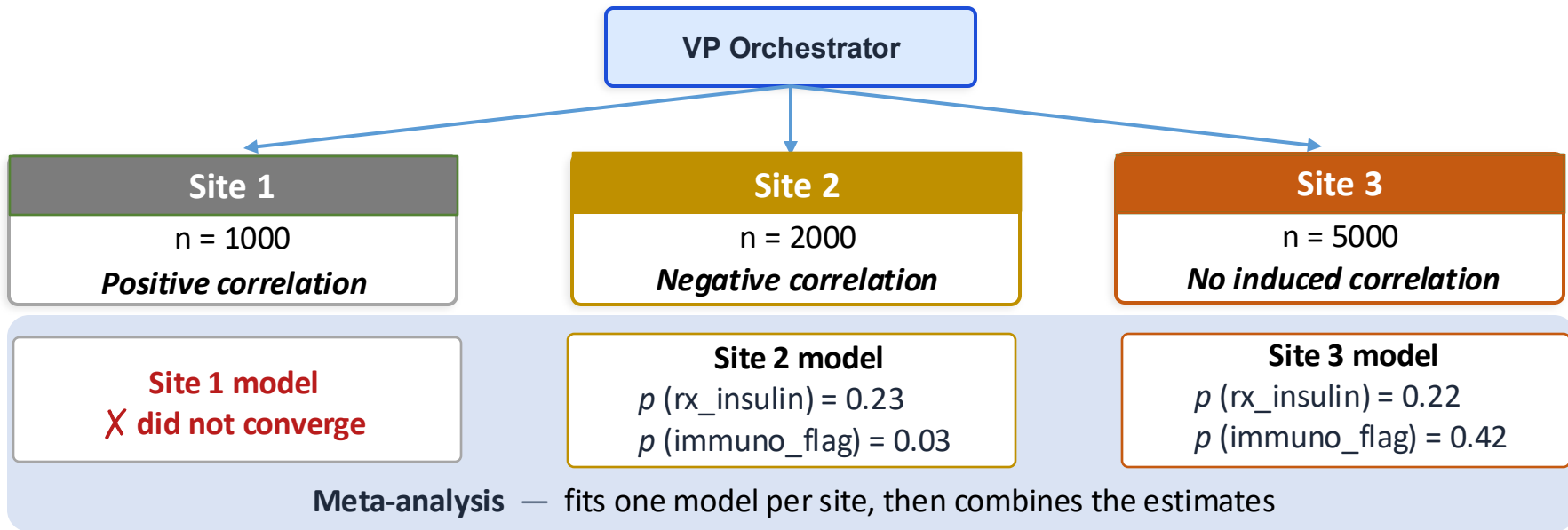
Results (VP vs Direct Pooling)

Cox co-efficient	51/51 identical*
p -value	51/51 identical*
Standard error	51/51 identical*

* Up to 6 decimal places

How robust is VP under skewed data distributions across sites?

Experiment setup: induced site-wise correlation between covariates *rx_insulin* (globally significant) and *immuno_flag* (not significant).



VP recovered exact estimates as direct pooling

$$p(\text{rx_insulin}) = 0.00$$
$$p(\text{immuno_flag}) = 0.13$$

Key takeaways from today's talk

- Existing mechanisms for multicenter studies struggle with **high governance, privacy risk, reduced accuracy, and sensitivity to skewed data distributions**.
- Virtual Pooling (VP) addresses this issue by combining a **federated analytics platform** with **accurate federated algorithms** for biostatistical operations.
- We validated VP by replicating a published N3C COVID-19 study (59 sites, ~300K patients), consisting of a workflow of propensity score matching followed by Cox survival analysis.
 - VP arrived at the **same matched cohort** and **exactly reproduced Cox coefficients, p-values, and standard errors** for all 51 covariates.
 - VP remained **robust to skewed site sizes** and to **opposing correlations between covariates**.

Thank you!