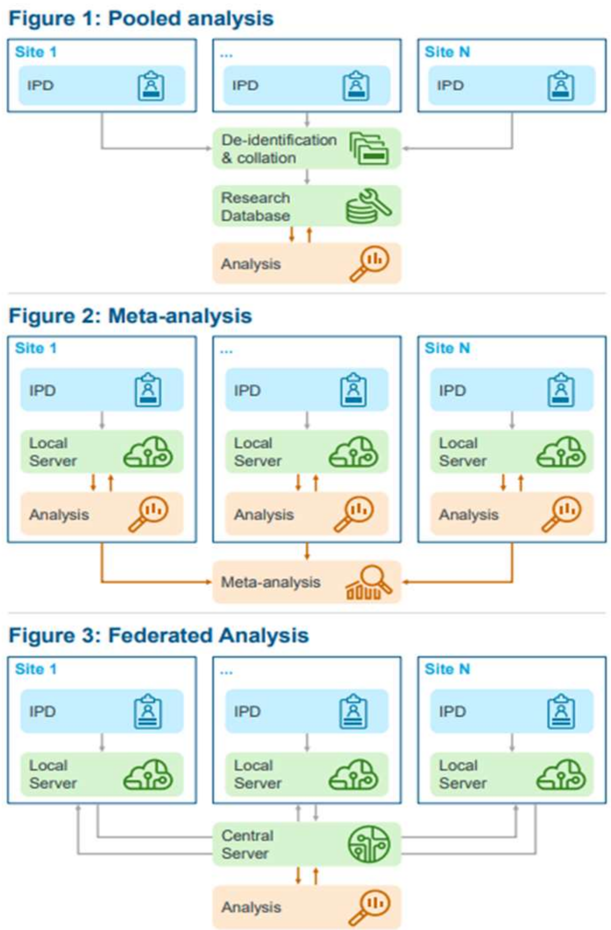
Collaborative acceleration for insights at scale

via a federated data and biosample network in Neuroscience

Dominik Heinzmann, Senior Director Data Science, Roche
Bjoern Tackenberg, Global Medical Science Group Leader, Neuroscience
Eric Boenert, Product Manager, Federated Open Science

Speaking the same language



Pooled analyses

- Used for statistical analyses and ML
- Sites give individual patient data (IPD) to an external party
  - Potential legal and trust issues

Meta-analysis

- Statistical analyses but not ML
- Only summary statistics obtained from IPD shared among institutions
- Many limitations, including ecological fallacy (eg Simpson’s paradox)

Federated analysis

- Statistical analysis and ML
- Results that are “equivalent” to pooled IPD
- Preserves privacy
  - **Federated Analytics** (FA): Applying statistical methods to the analysis of “raw” data that is stored locally at multiple institutions (and remains there)
  - **Federated Learning** (FL): Training a machine learning model across multiple institutions without centralizing data

Today's Objectives

Introduction of the federated data and biosample network

- What
- Why
- How

Core building blocks

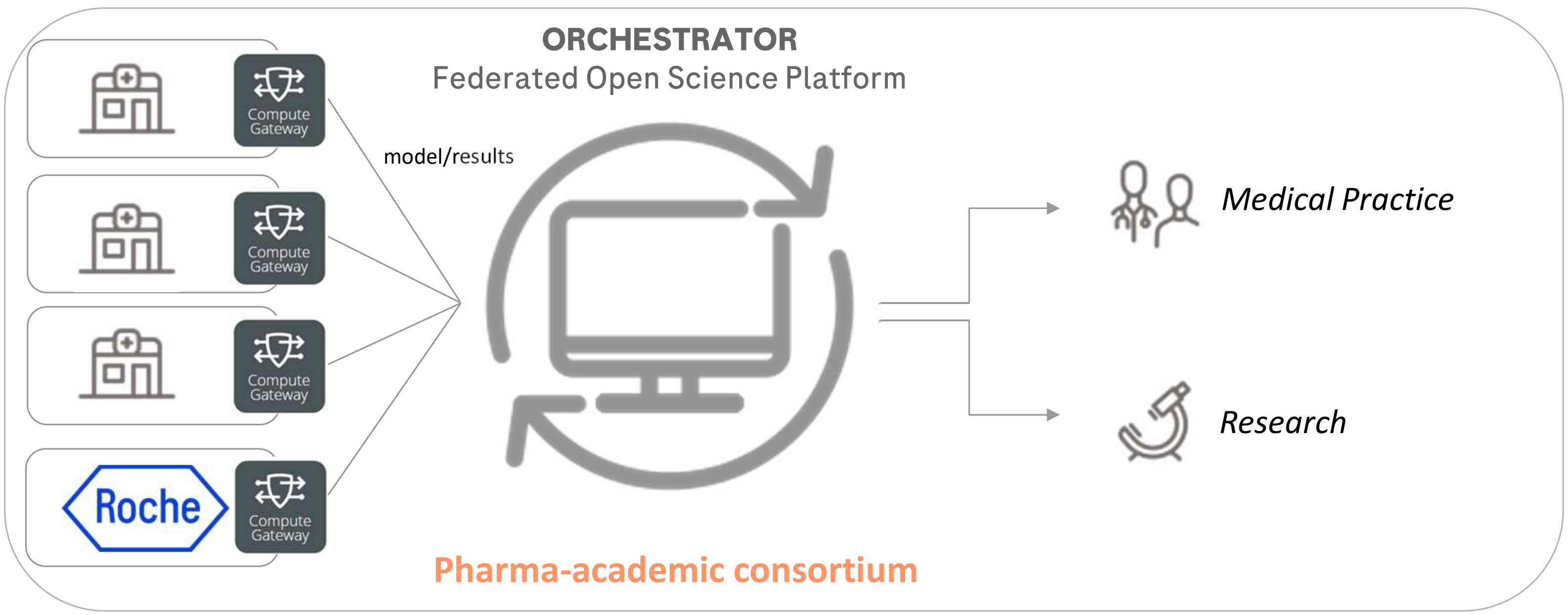
- Data
- IT
- Statistics / ML

WHAT

Novel ways of collaboration in partnership

- **MORE DATA by securely enabling more diverse patient pool and more modality of data**
  - Across countries, privacy-preserving access to patients & more data modalities (imaging, m-omics…))
  - Data stays “at home”, algorithm is sent to “home” and aggregated results are sent back and “federated”

- **MORE INNOVATION by scaling “collaboration in partnership”**
  - extending the currently applied approach with fewer familiar & like-minded & highly trusted relationships into broader collaborations (*diversity & inclusion*)
  - reducing the risk of cognitive lock-ins and hence increasing innovation power



WHY

Value Proposition of a federated data & biosample network (Neuroscience)

Value for Patients & Society

Patient	Science	Society
▪ **Precision neuroscience enablement** through access to high quality data and biosamples of >>10’000 patients per disease	▪ **New target** and **biomarker discovery** ▪ **Innovative endpoint** development (e.g. vocal biomarkers as surrogate endpoint for disability in Multiple Sclerosis)	▪ Platform to **validate and measure societal value** and **impact** of therapeutics & integrated healthcare solutions

Value for Roche

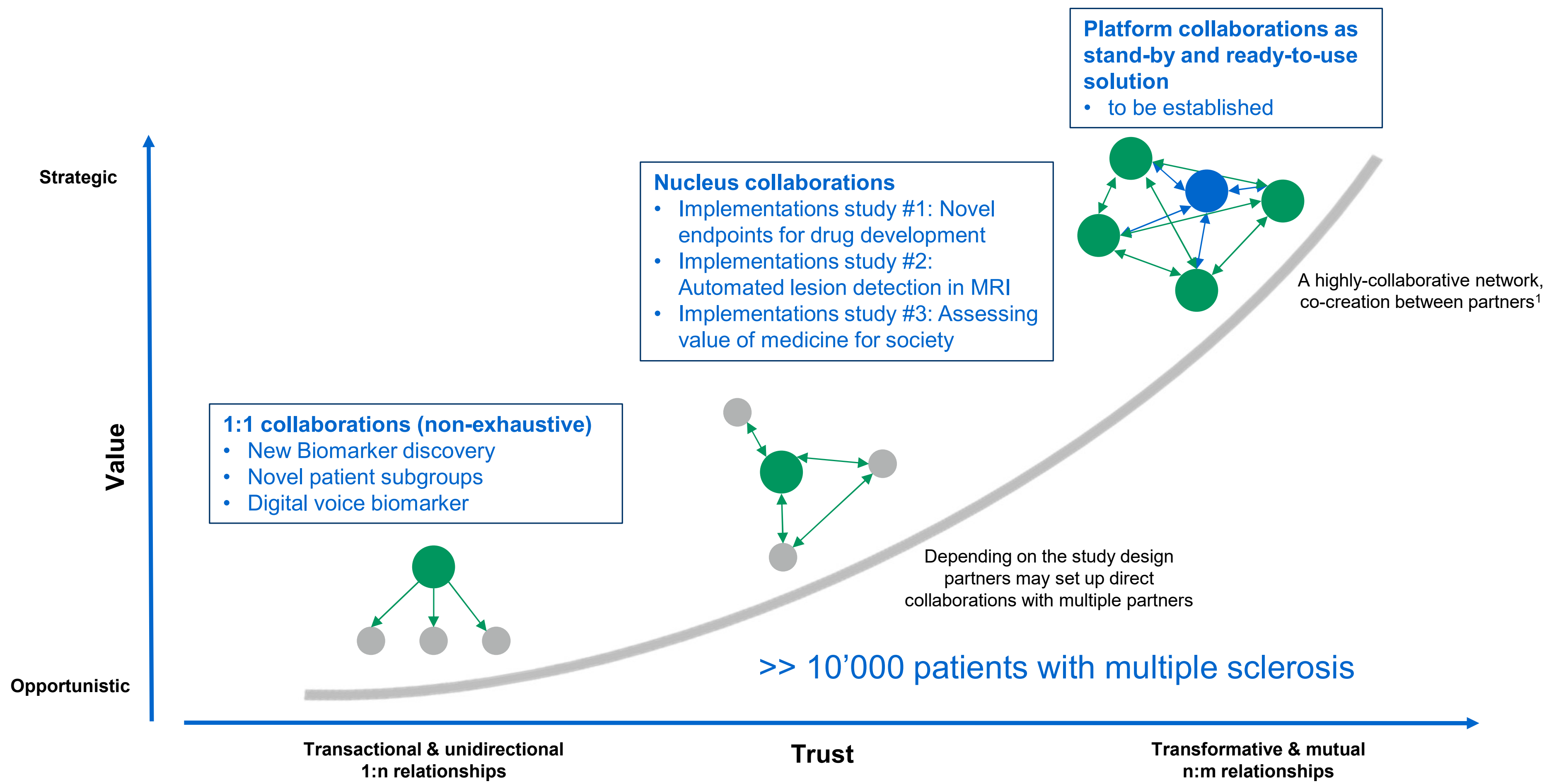
DISCOVERY	DEVELOPMENT	SHAPING MEDICAL PRACTICE
New target discovery	Novel endpoint development	Precision neuroscience enablement
		Societal value investigation
New biomarker / dig. measurements / imaging marker identification	Feasibility studies for fast decision making	Clinical decision support Development

ULTIMATIVELY: Enable next generation evidence-based medicine* in Neuroscience (and beyond)

*Subbiah (2023), “The next generation of evidence-based medicine”, Nature Medicine

HOW

Scope of collaboration shifting from narrower to broader (“crowds”)



[1] according to Yaqub M. Z. et al. (2020), “Network innovation versus innovation through networks”, Industrial Marketing Management

Today's Objectives

Introduction of the federated data and biosample network

- What
- Why
- How

Core building blocks

- Data
- IT
- Statistics / ML

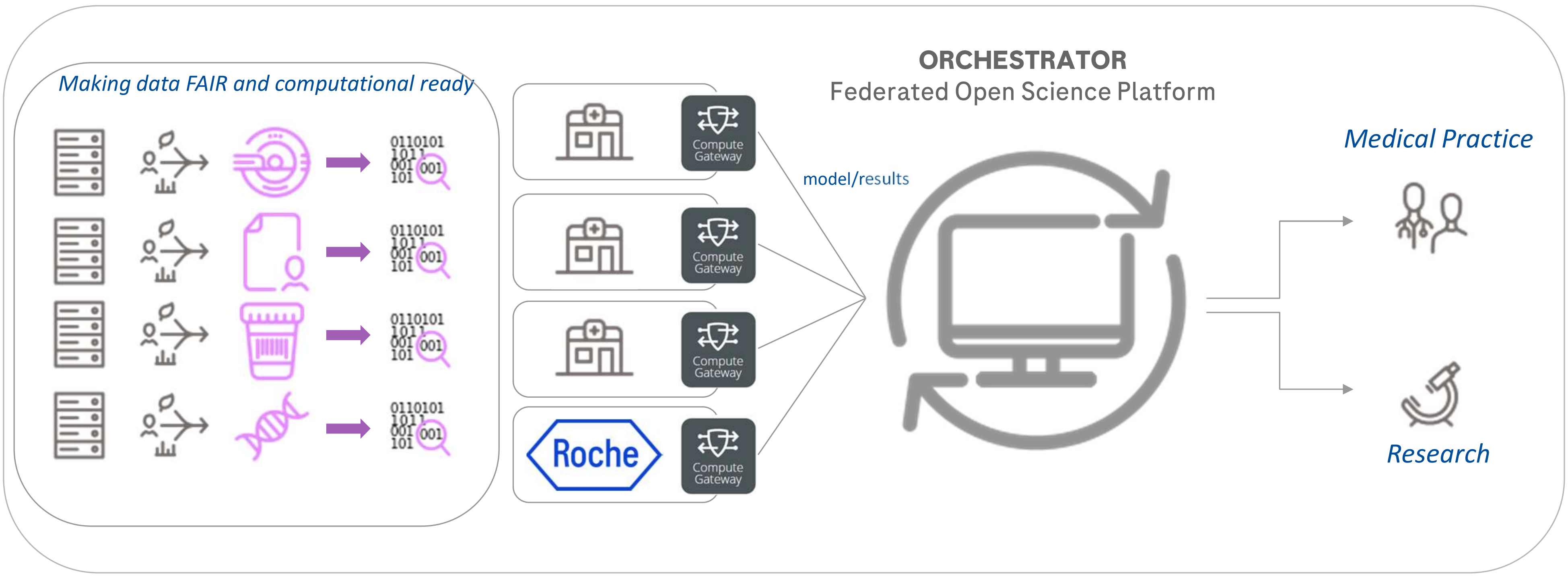
Core building blocks: Data, IT and Statistics / ML

Fast and secure access to more multi-modal data

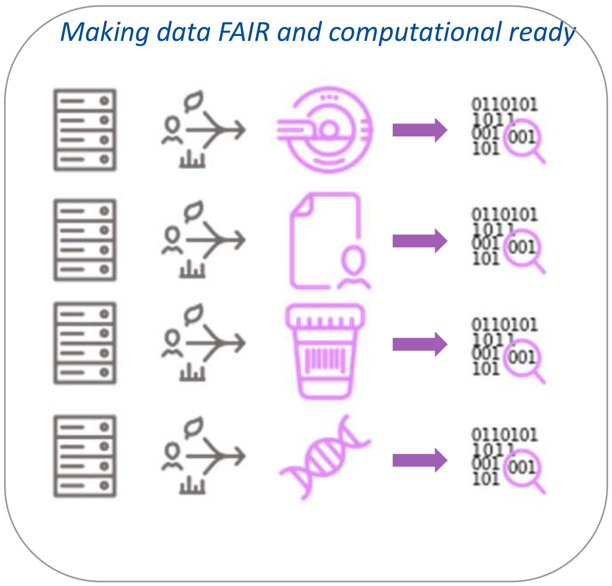
Governance: Control over data

Patient privacy-preserving

Fast turn-around on statistical analyses & ML



Data: Substantial investments needed into pipeline



Data pipeline at the "nodes"

- Often, suboptimal data pipeline in universities and hospitals for multi-modal data (e.g. different databases for different modalities)
- Substantial support (financial and/or technical) needed for ETL (Extract, Transform, Load), i.e. combine multiple sources into a data warehouse at university/hospital site
- Need dedicated university/hospital resources for data management, data science, disease expertise (MD) and IT next

Data model

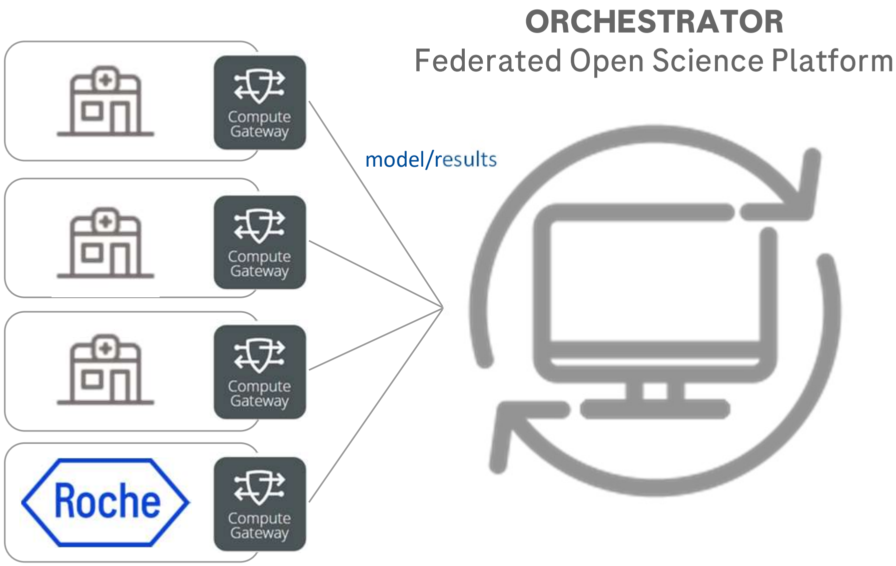
- Prioritization: Research questions now versus future ones
- Generic versus tailored core data models (CDM)
    - Challenge: RWD versus clinical trial data
- Analysis-ready datasets (ARD)
    - Pre-processing on hospital level versus pre-processing as part of computation via federated platform (privacy and bias considerations)

"Real-time" data catalogue including metrics on data quality

- As data pipeline is in flow, up to date inventory needed
- For data quality, see also BBS webinar March 2023 on RWD quality*

*https://baselbiometrics.github.io/home/docs/events_past.html

IT: Privacy, Security and Goverance



A secure and privacy-preserving infrastructure must **empower collaborators** to make sensitive data available by maximizing the degree of control.

Some key characteristics:

- Ability to review, add and remove the availability of any given dataset for a federated project
- Ability to approve, audit and monitor the execution of a specific federated workflow
- Ability to support output privacy e.g. secure aggregation
- Ability to review, audit, and customize the node/client deployment
- Ability to set fine-grained controls in a user-friendly way
- Ability to detect disclosive statistical computations
- Ability to apply identity and access management, network segmentation, secure network communication, external attack protection and others

Roche's current threat model assumes that nodes and servers are **honest-but-curious**, nodes are independent actors and do not collude.

Statistics: Federated analytics (FA)

- **OVERVIEW: Quite active recent development of new algorithms for privacy preserving FA**
- **EXAMPLE:** Cox proportional hazard model

  - **WebDISCO**[1] (horizontally partitioned, R-based)
    - Iterative estimation
    - Based on Newton-Raphson approach to estimate parameter for Breslow's partial likelihood function of cox model
    - At each iteration, aggregates gradient & Hessian from all sites
    - **Challenge:** Large communication costs

  - **Andreux et al**[2] (horizontally partitioned, Python-based)
    - Iterative estimation
    - Based on discrete-time extension of cox PH model to formulate survival analysis as a classification problem with separable loss function
    - Idea: Binning of observed times into finite set, where then hazard = conditional probability rather than a rate
    - Close to WebDISCO[1] algo, but optimizes communication costs, may provide better privacy preserving characteristics and allows more flexible tackling for non-linear relationships (e.g. neural networks)

[1]Lu, Wang, Ji et al. (2015), "WebDISCO: a web service for distributed cox model learning without patient-level data sharing", J Am Med Inform Assoc,
[2]Andreux, Manoel, Menuet et al. (2020), "Federated survival analysis with discrete-time Cox models", arXiv preprint arXiv:2006.08997

Statistics: Federated analytics (FA)

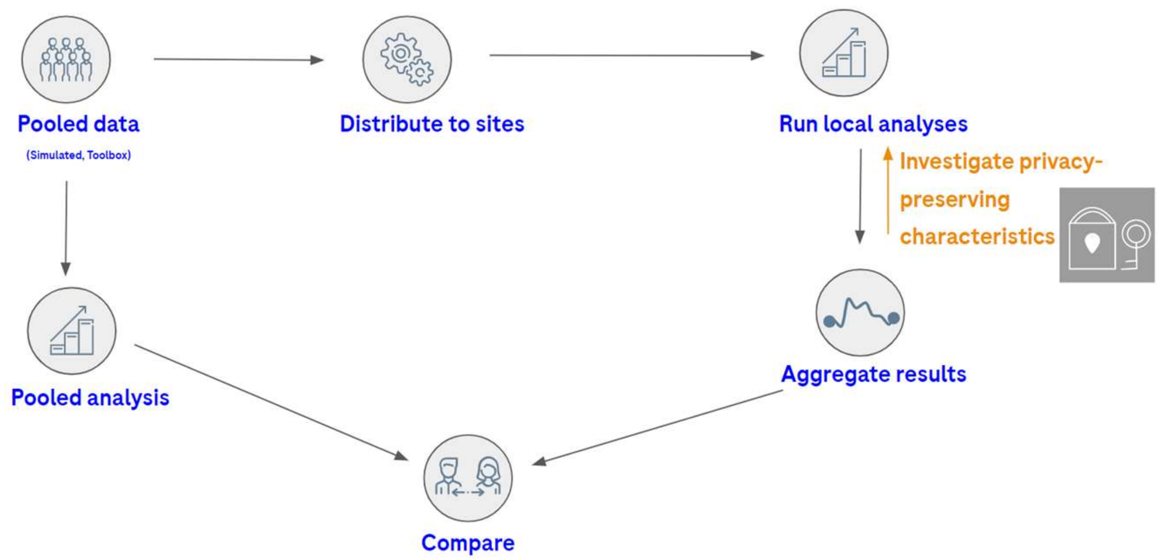
- **EXAMPLE (cont'd):** Cox proportional hazard model
  - **Yu et al**[3] (horizontally partitioned, code N/A)
    - Non-iterative estimation
    - Approximates cox model on dimensionality-reduced data (linear transformation)
    - Idea: Lower-dimensional projections are in general not reversible (but…)
    - Instead of $\{\mathbf{x}_i, \mathbf{t}_i, \boldsymbol{\delta}_i\}$ share $\{\mathbf{B}^T \mathbf{x}_i, \mathbf{t}_i, \boldsymbol{\delta}_i\}$ (**B** the same for all sites), aggregate and learn cox model
    - Optimization task: Preserve relationship of data points
    - **Challenge:**
      - Designed for making predictions, does not estimate target model parameters
  - **DC-COX**[4] (horizontal & vertical, Python-based)
    - Non-iterative estimation
    - Each node *individually* constructs a dimensionality-reduced representation and share with orchestrator
    - But apply it to site data & an anchor dataset (same for all sites) and share back to orchestrator
      - $\widetilde{\boldsymbol{X}}_i^{anchor}$ = Projection($\boldsymbol{X}_i^{anchor}$), $\widetilde{\boldsymbol{X}}_i^{original}$ = Projection($\boldsymbol{X}_i^{original}$)
    - Orchestrator generates collaboration presentations such that for all sites i, k:
      - $\widetilde{\boldsymbol{X}}_i^{anchor}\mathbf{G}_i \approx \widetilde{\boldsymbol{X}}_k^{anchor}\mathbf{G}_k$
      - Apply $\mathbf{G}_j$ to $\widetilde{\boldsymbol{X}}_j^{original}$ for all sites j, and apply Cox PH to resulting data & share back

[3] Yu, Fung, Rosales et al. (2008), "Privacy-preserving cox regression for survival analysis", Proceedings ACM SIGKDD Aug 2008
[4] Imakura, Tsunoda, Kagawa et al. (2023), "DC-COX: Data collaboration Cox proportional hazards model for privacy-preserving survival analysis on multiple parties", Journal of Biomedical Informatics

Statistics: Federated analytics (FA)

- **Previous methods can be combined with other privacy-enhancing approaches (differential privacy, AI-generated synthetic data, homomorphic encryption…)**
  - Privacy versus utility balance

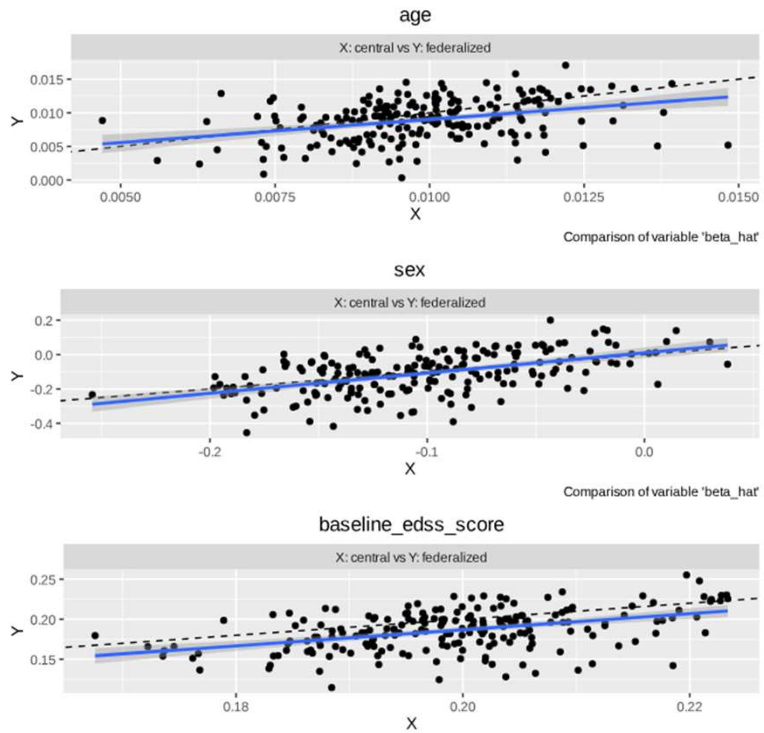
- **Importance to properly developing & validating algorithms**
  - Dry-run of the platform with using simulated/synthetic data (to know the ground truth)
    - Part 1: Virtual nodes (quick iterations, but challenging to test e.g. site infrastructure performance)
    - Part 2: Full network site nodes but simulated/synthetic data



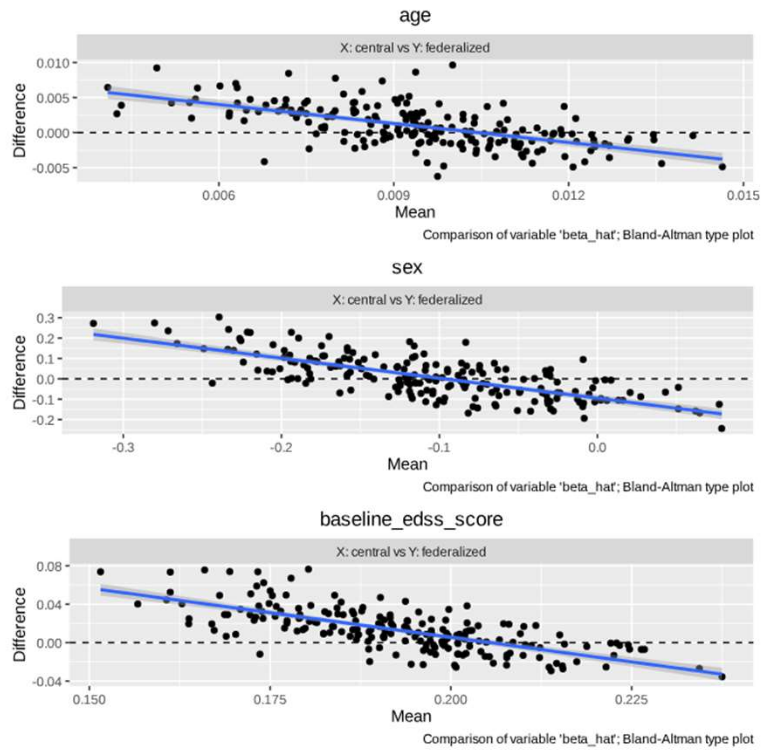
Statistics: Federated analytics (FA)

Initial results ***- work in progress***
DC-COX with marginally heterogeneity among sites

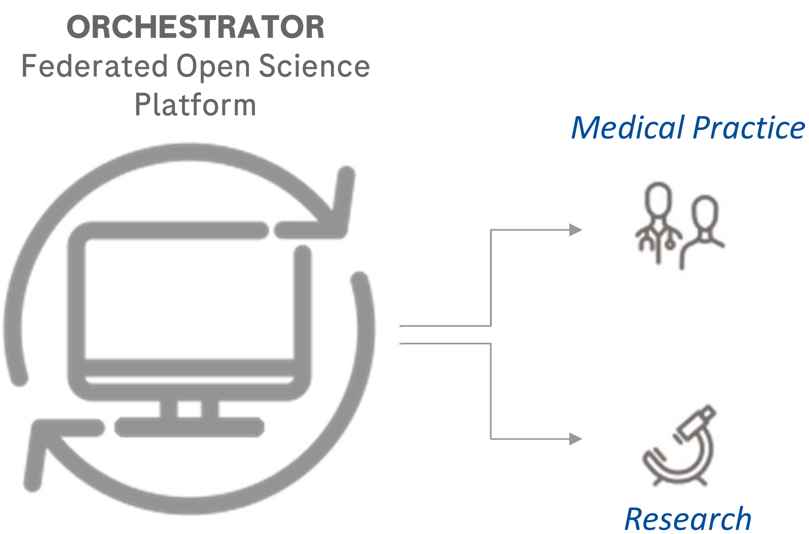
Correlation plots



Bland-Altman plots



Statistics: Fast turn-around on analyses (FA)



In build phase: Challenges to prepare Statistical Analysis Plan (SAP) in absence of “knowing” data structure and completeness

- **Key problem** in the initial phase of a project (where you still build CDM) as over time, knowledge on data structure, quality etc in the network will grow
- **Example challenge:**
  - Mot many *off the shelf* solution available
  - Strategies for handling missing data remains a major bottleneck in real-world FL/FA deployment
  - Typically performed locally, but is likely biased, since the subpopulations locally observed at the hospitals may not be representative of the overall one
  - Federated versions are evolving, but need some time for proper implementation and validation

Discussion

- We are in the process of implementing and scale a federated data and biosample network for data on patients with multiple sclerosis
- We learn and iterate the network to support next-generation evidence generation*
- Still significant method gap limiting broad adoption of FA in real-world studies
  - Imputation approaches (MI,…)
  - causal inference (marginal structural models, G-estimation…)
  - hierarchical models

- **Essentials for success**
  - A close collaboration “in partnership” between academics, hospitals and pharma are needed to unleash the full potential of such a network
  - A well orchestrated teamwork with key players from data management, data science/statistics, medical doctors and IT
  - Developing of federated analytics functions challenging as one need to consider many dimensions including privacy, speed (as a function also on computing power across nodes & at orchestrator)

*Subbiah (2023), “The next generation of evidence-based medicine”, Nature Medicine

Federated data & biosample network

accelerating - catalyzing – shaping

Doing now what patients need next