Pseudonymisation in medical research: from theory to practice

Pseudonymous data: processing personal data while mitigating risks

Prof. Dr. Fabian Prasser
Hammam Abu Attieh, M.Sc.
Armin Müller, M.Sc.

08.12.2021

BIH Berlin Institute of Health @Charité

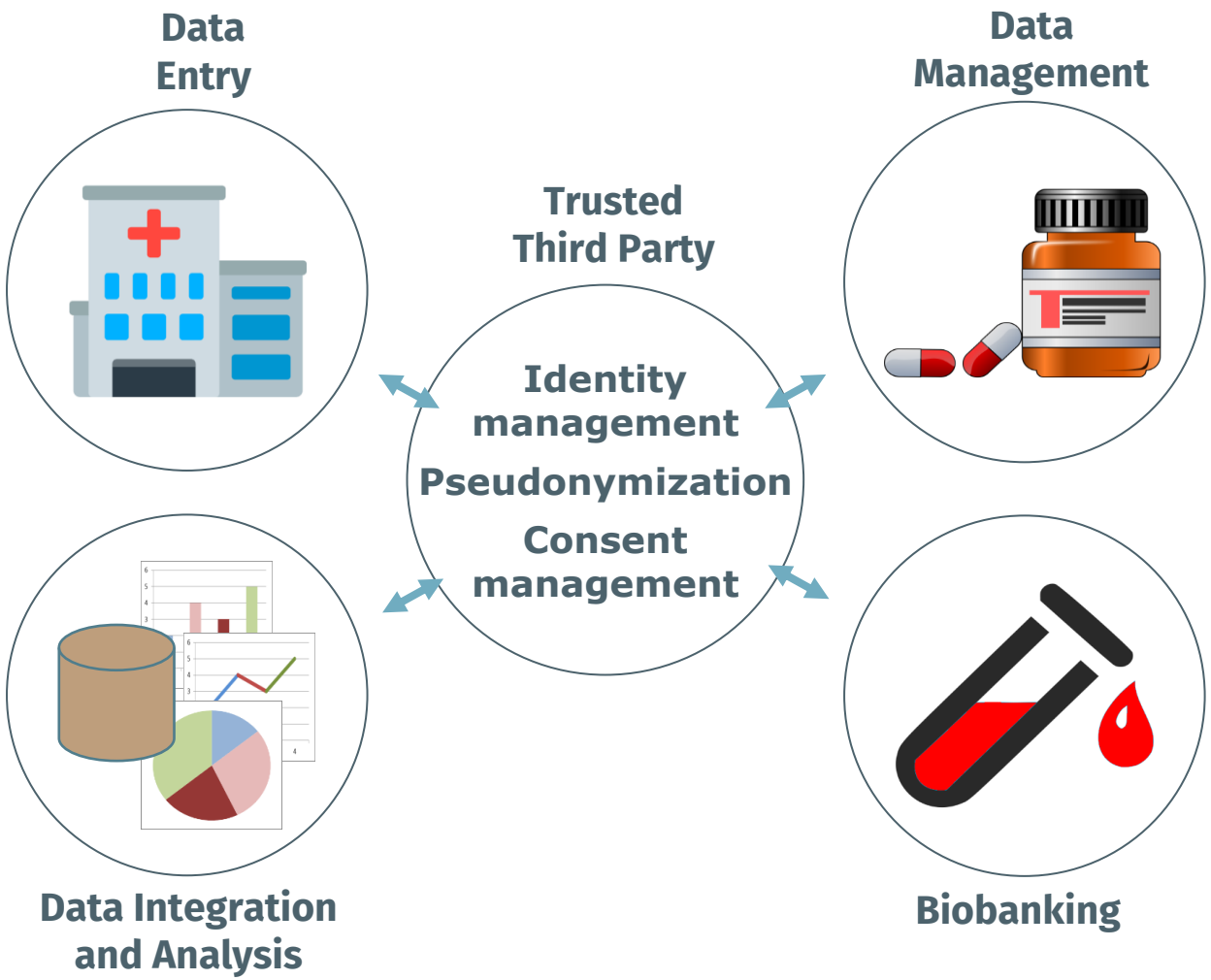
Motivation: Legal Requirements

- **Data concerning health (in the sense of Art. 4 No. 15 GDPR):**
  - "'Data concerning health' means personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status."

- **Art. 9 GDPR "Processing of special categories of personal data"**
  - "data concerning health"

- **Pseudonymisation (in the sense of Art. 4 No. 5 GDPR):**
  - "'Pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information [...]"

Motivation: Medical Research Areas



Source: TMV e.V. Drepper, J., Semler, S.

Different areas of medical research

- are subject to different legal frameworks
- therefore need different ethics and data protection concepts

Data must often remain identifiable

- for consultative participation (e.g. by experts)
- for correct assignment of longitudinal data
- for quality management
- for re-contacting in case of incidental findings

Heterogeneity of Pseudonymisation in the Literature

Demiroglu et al. (2012): Large-scale research project in the area of psychiatric genetics

Aamot et al. (2013): Sample and data management in translational research for oncology patients

Pommerening et al. (2006): Infrastructure for longitudinal studies involving medical data, genetic data and data for managing collections of biosamples

Neubauer et al. (2011): Approach in which smart cards allow patients to control the de-pseudonymization process

Gulcher et al. (2000): System for collecting research data and biosamples for disease-based gene discovery projects

Spitzer et al. (2009): Approach that utilizes pseudonymization to secure a web-based teleradiology platform for exchanging digital images

Authenticity	Integrity	Accountability	Confidentiality	Availability	Authorization
X	X		X	X	X
X	X	X	X		
X	X	X	X		
X	X		X	X	X
		X	X		
X	X		X		X

Source: Kohlmayer F, Ronald L and Prasser F. "Pseudonymization for research data collection: is the juice worth the squeeze?." *BMC Medical Informatics and Decision Making* 19.1 (2019): 1-7.

Heterogeneity of Application Scenarios: ENISA Guideline

Pseudonymization scenarios from the ENISA guideline "Pseudonymisation techniques and best practices - Recommendations on shaping technology according to data protection and privacy provisions"

- **Scenario 1 - Pseudonymization for internal use**: „[...] when data are collected directly from the data subjects and pseudonymised by the data controller, for subsequent internal processing."
- **Scenario 2 – Processor involved in pseudonymization:** „[...] variation of scenario 1, where a data processor is also involved in the process by obtaining the identifiers from the data subjects [...]"
- **Scenario 3 - Sending pseudonymized data to a processor**: „[...] the data controller again performs the pseudonymisation but this time the processor is not involved in the process but only receives the pseudonymised data from the controller."
- **Scenario 4 - Processor as pseudonymization entity**: „[...] the case where the task of pseudonymisation is assigned by the controller to a data processor (e.g. a cloud service provider that manages the pseudonymisation secret and/or arranges the relevant technical facilities)."
- **Scenario 5 - Third party as pseudonymization entity**: "[...] the pseudonymisation is performed by a third party (not a processor) who subsequently forwards the data to the controller. Contrary to the Scenario 4, the controller in this scenario does not have access to the data subjects' identifiers (as the third party is not under the control of the data controller).

Pseudonymization in Practice: NUM CODEX

Funded by
Federal Ministry of Education and Research

NUM: Netzwerk Universitätsmedizin (Network of University Medicine)

- Large-scale research network of 36 university hospitals in Germany introduced as a response to the COVID-19 pandemic

netzwerk universitäts medizin

MII: Medizininformatik-Initiative (Medical Informatics Initiative)

- Large-scale national project to establish a data sharing infrastructure for access to healthcare data for secondary purposes

MEDIZIN INFORMATIK INITIATIVE

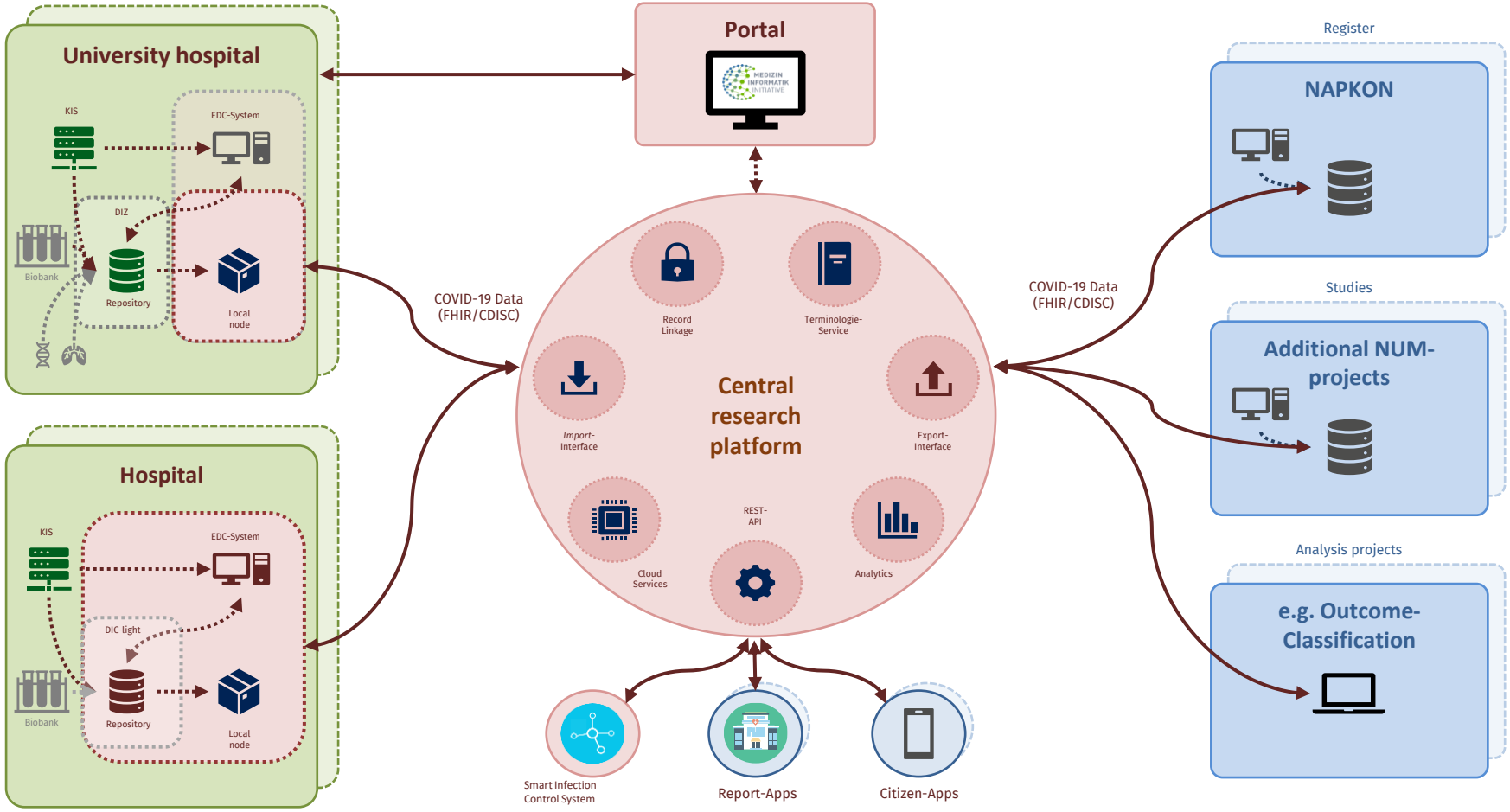
CODEX: Covid-19 Data Exchange Platform

- Platform for access to real-world data on COVID-19 patients established by the MII under the umbrella of NUM

CODEX

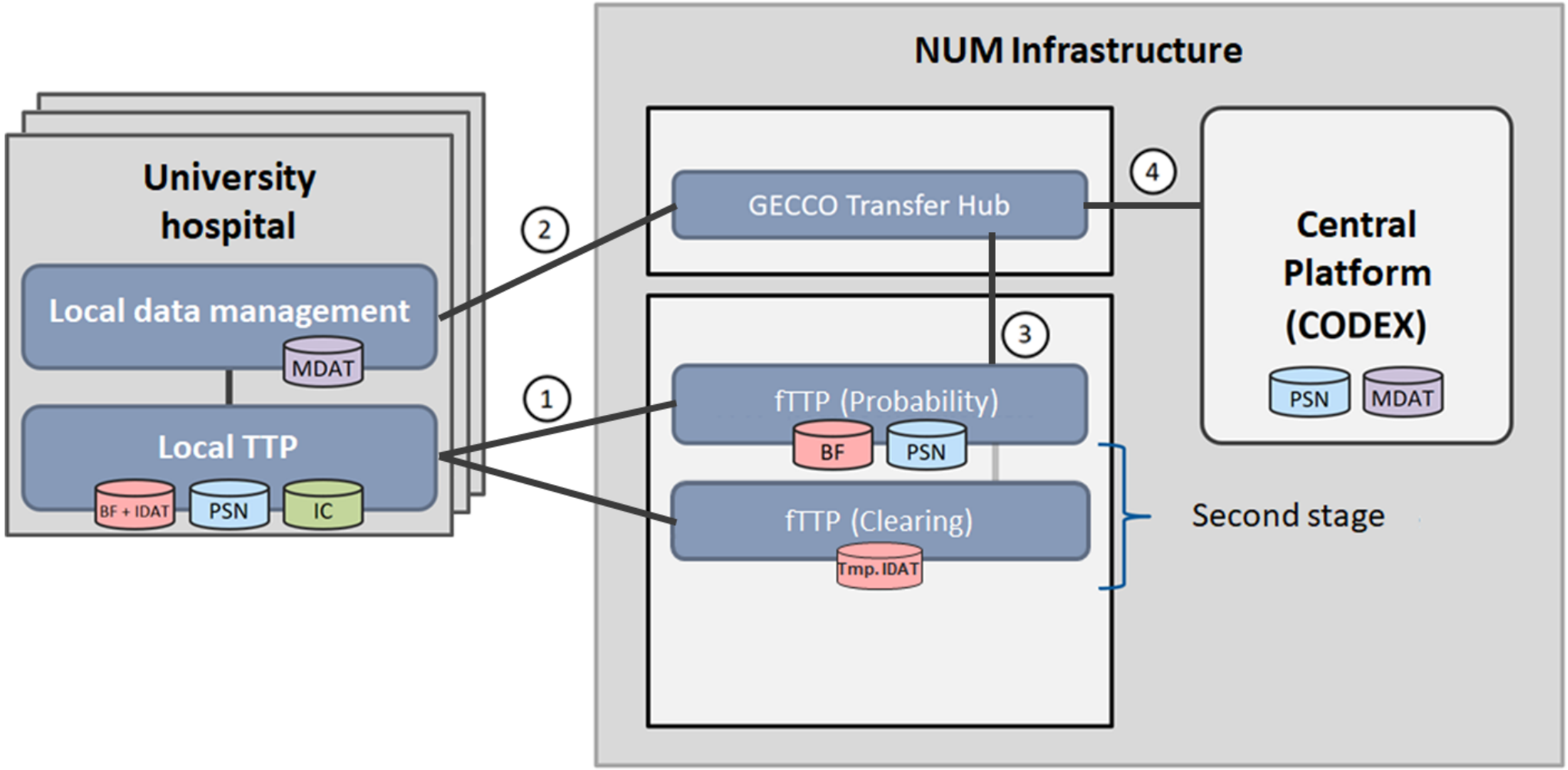
BIH Berlin Institute of Health @Charité

Pseudonymization in Practice: CODEX Architecture



Source: Netzwerk Universitätsmedizin. Umsetzungskonzept für den Aufbau einer nationalen Forschungsdateninfrastruktur der Universitätsmedizin zu Covid-19. Version 1.05.

Pseudonymization in Practice: CODEX Architecture



Modified from: T. Bahls, M. Bialke, W. Hoffmann, C. Fegeler, H. Hund

Pseudonymization in Practice: CODEX Architecture

Four different pseudonymization points

Pseudonymization for internal use (ENISA Scenario 1)

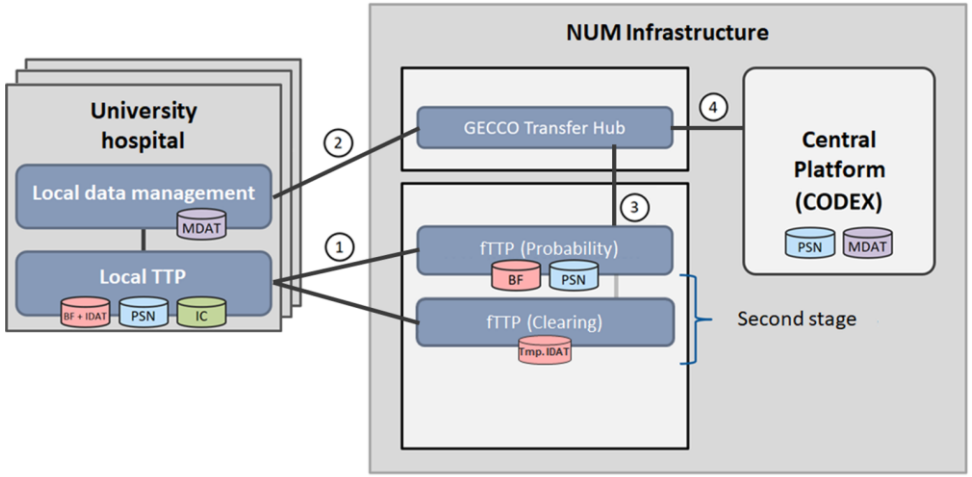
1. Local pseudonymization in the university hospitals

Sending pseudonymized data to a processor (ENISA Scenario 3)

2. Pseudonymization through the fTTP Probability
3. Pseudonyms through the fTTP Clearing

Third party as pseudonymization entity (ENISA Scenario 5)

4. Pseudonyms when researchers export data from the central platform



BIH Berlin Institute of Health @Charité

Pseudonymization in Practice: ORCHESTRA

Goal

- Establishing an European large-scale cohort for retrospective and prospective studies on the prevention and treatment of COVID-19
- Rapid scale-up to generate evidence at the speed needed in times of a pandemic

Network

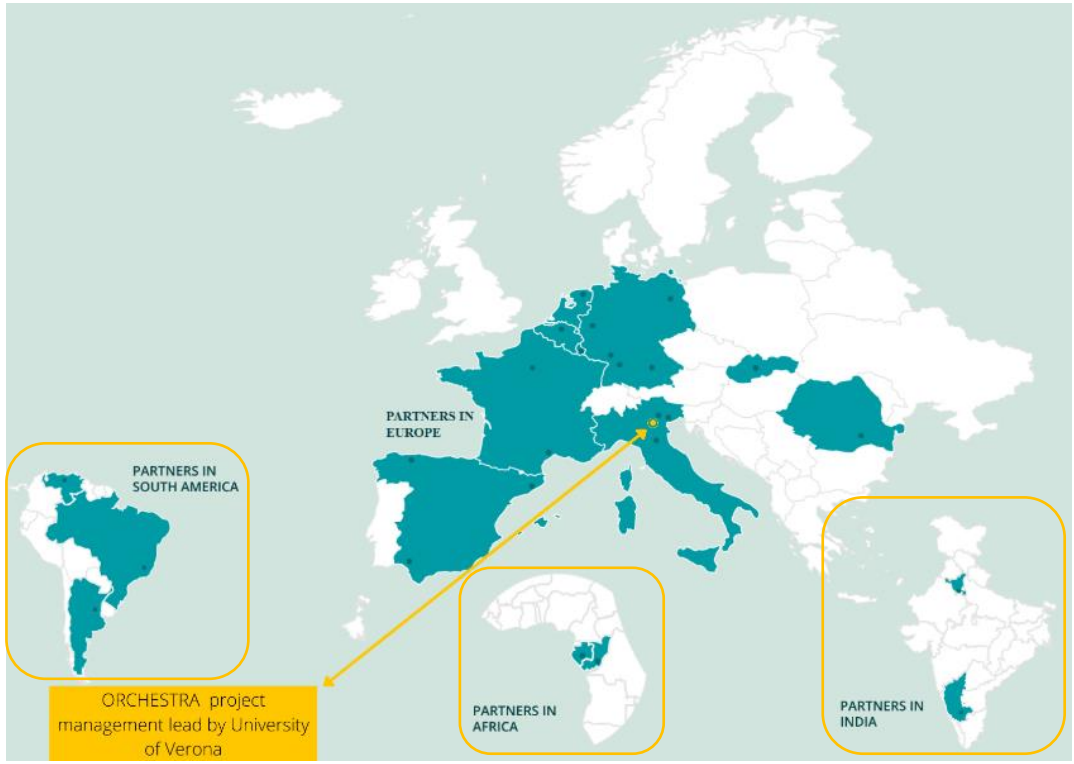
- 26 partners (extending to a wider network of 37 partners) from 15 countries from all over the world

Data

- Data and samples from up to 1 million patients

European Commission | Horizon 2020 European Union funding for Research & Innovation

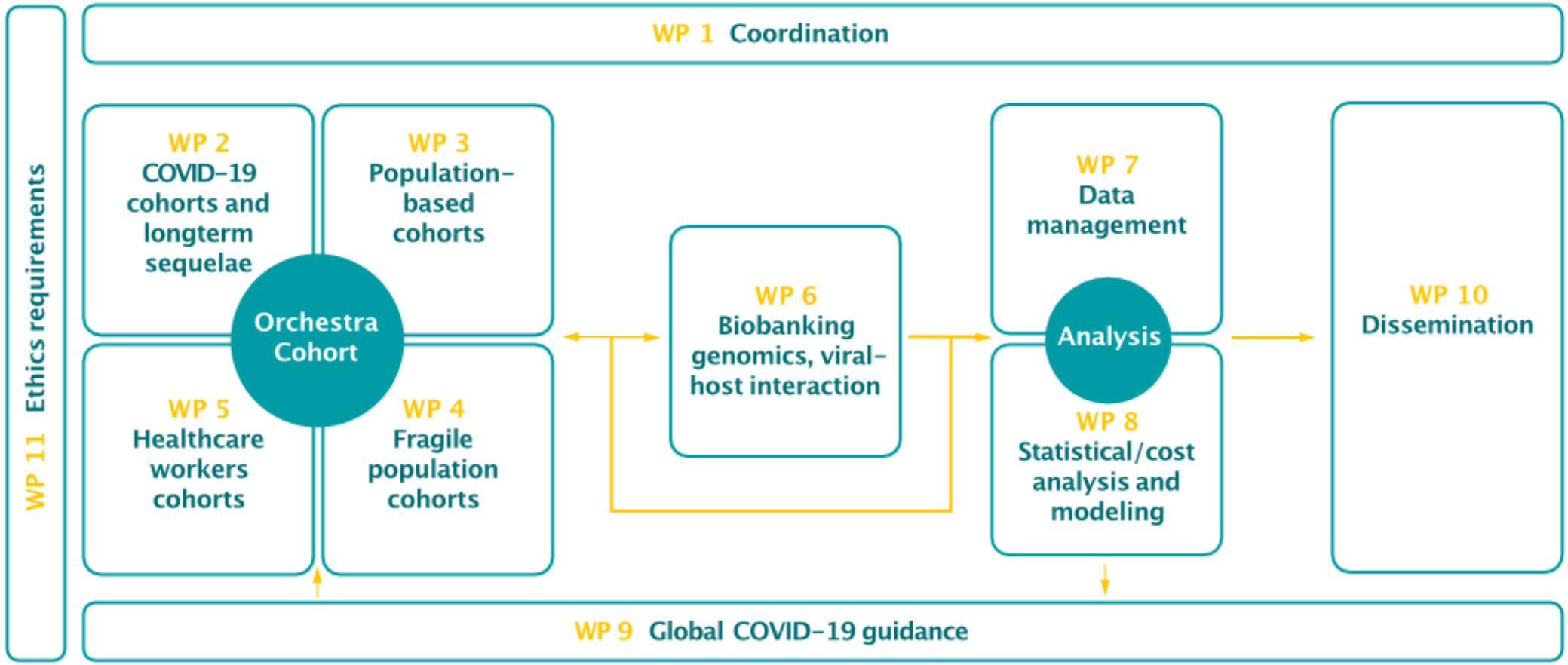
ORCHESTRA



Source: https://orchestra-cohort.eu/partners/

BIH Berlin Institute of Health @Charité

Pseudonymization in Practice: ORCHESTRA Cohorts and WPs



Source: https://orchestra-cohort.eu/work-packages/

Pseudonymization in Practice: ORCHESTRA Challenges

Heterogeneous legal framework

- EU (GDPR & country-specific regulations) and non-EU law

Heterogeneous technical infrastructures

- Different OS, different local network structures and expertise, different degrees of internet connectivity

Heterogeneous study types

- Prospective, retrospective

Heterogeneous datasets

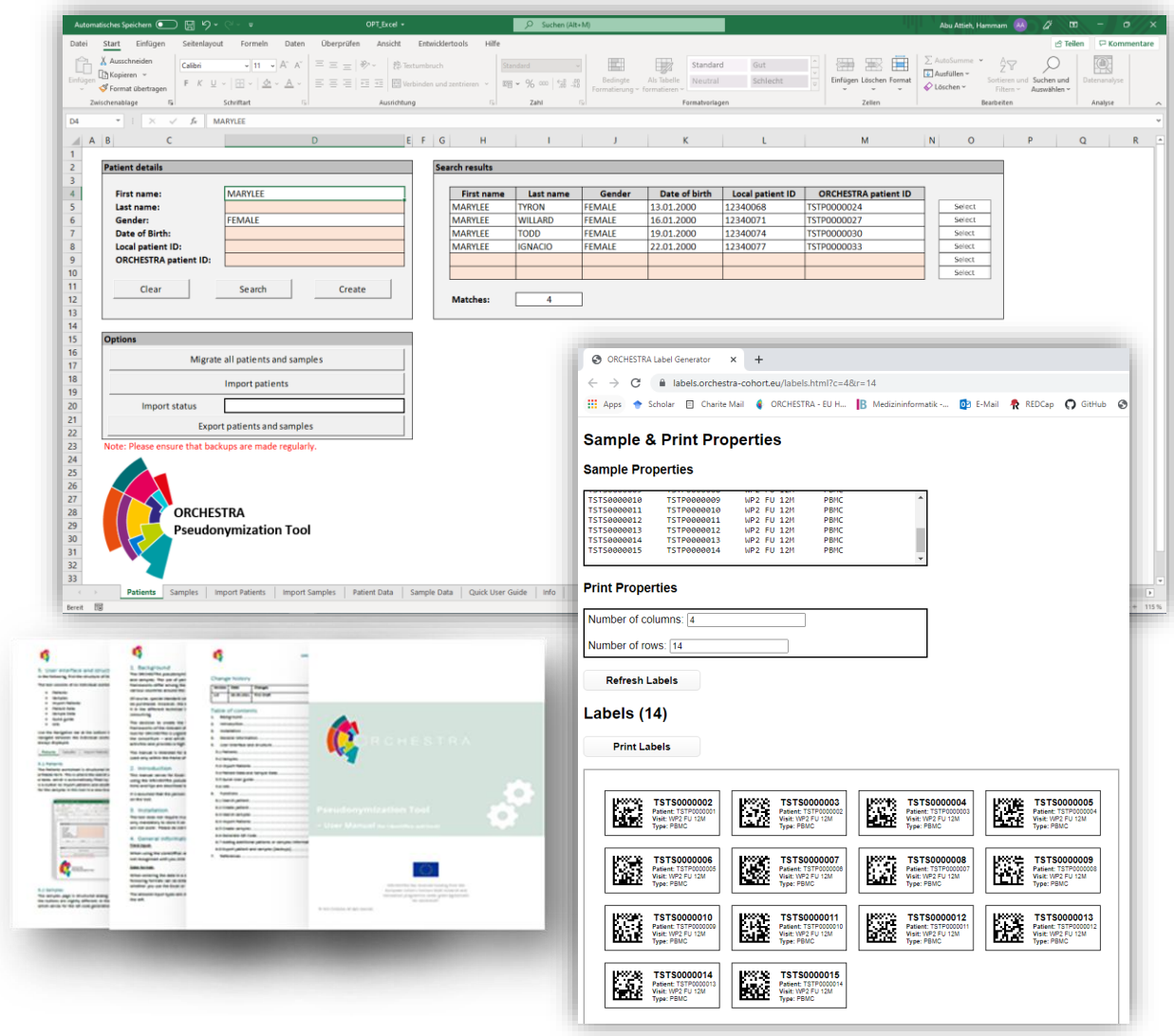
- Dynamic cohorts (e.g. 3rd dose), different target groups (vaccinated, infected, etc.)

→ Challenging environment to roll-out pseudonymization methods

→ Tight deadlines

Pseudonymization in Practice: OPT

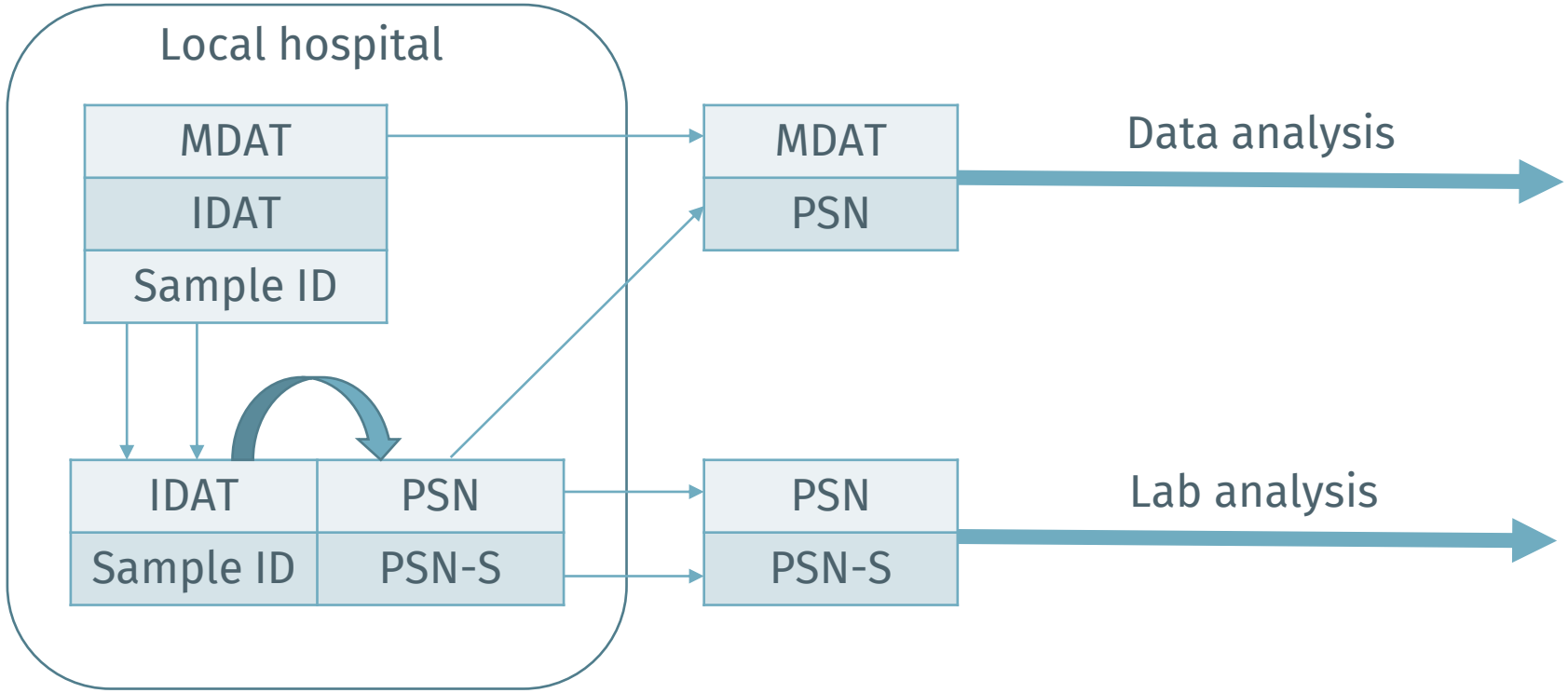
- The ORCHESTRA Pseudonymization Tool is based on a “runtime environment” that is basically available everywhere: spreadsheet software
  - No installation of server software needed
  - Special caution is needed → Handbook
- Supports the management and pseudonymisation of patient and sample identifiers
- Supports local record linkage
- Is integrated with laboratory printers to print labels for tubes



BIH Berlin Institute of Health @Charité

Pseudonymization in Practice: OPT Usage Scenario

IDAT – Identifying Data
MDAT – Medical Data
Sample ID – Local sample ID
PSN – Pseudonym for IDAT
PSN-S – Pseudonym for sample ID



Pseudonymization scenario*: Sending pseudonymized data to a processor.

* as in the European Union Agency for Cybersecurity (ENISA) guide "Pseudonymisation techniques and best practices - Recommendations on shaping technology according to data protection and privacy provisions (Nov. 2019)"

Discussion: Pseudonymization for Data Entry

- Pseudonymisation can significantly complicate the implementation of measures to maintain the rights of data subjects
- Authorized re-identification of patients in research data collection systems is a common procedure
- Implementation of suitable systems is complex
- Lack of flexible tools that can be used quickly and easily

- Technical and organizational separation of systems managing identified data, pseudonyms and research data leads to drawbacks
  - Increased number of interfaces
  - Maintenance challenges

→ Complexity is the enemy of security? *Geer DE. Complexity is the enemy. IEEE Security & Privacy. 2008;6(6):88–8.

Thank you for your attention!

Prof. Dr. Fabian Prasser
Medical Informatics Group
BIH @ Charité - Universitätsmedizin Berlin

Contact:
mi.bihealth.org
fabian.prasser@charite.de

BIH Berlin Institute of Health @Charité