Data access and integration in a distributed computational environment for medical research

Piotr Nowakowski

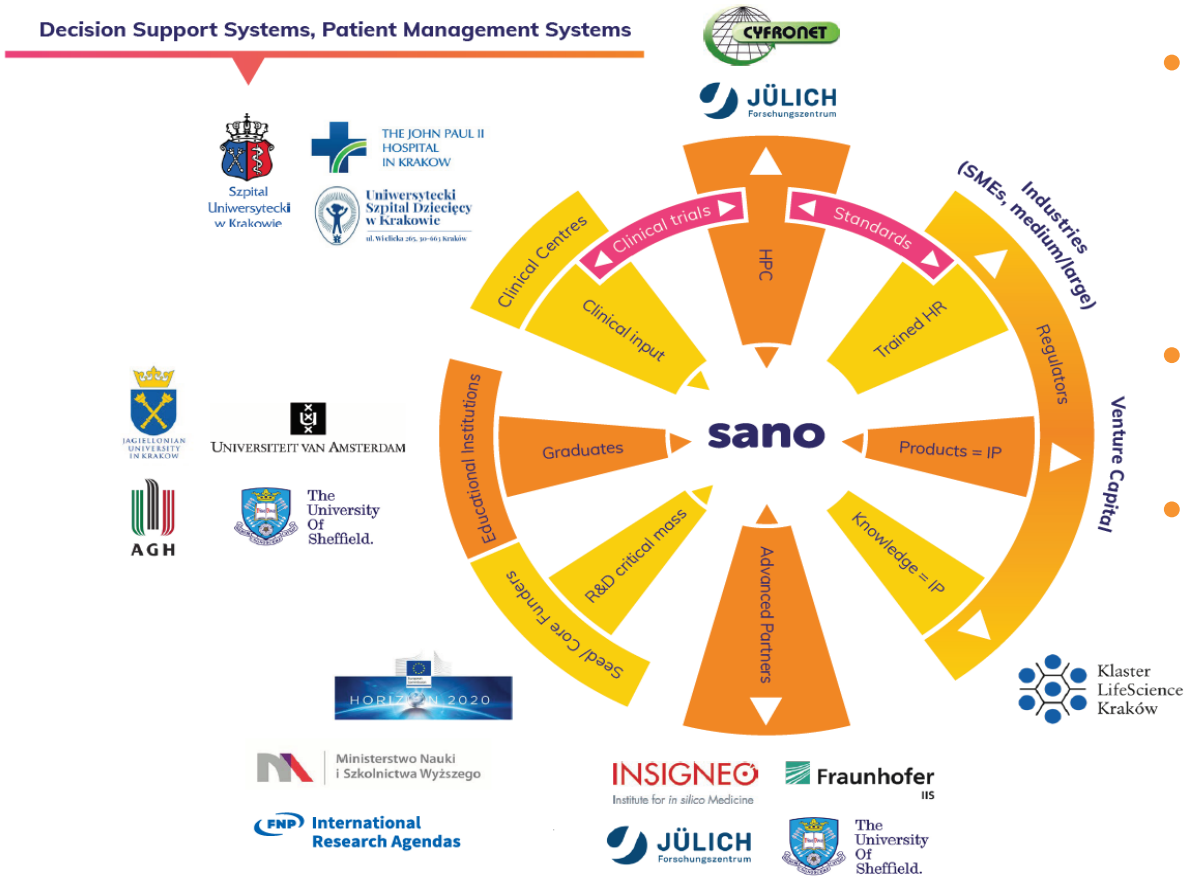
Sano Centre for Personalised Computational Medicine

p.nowakowski@sano.science

Overview

- *Introduction to the mission of Sano*
- *The Sano computational toolkit – tools and services*
- *Research initiatives*
- *Data access challenges*
- *Closing remarks*

Sano at a glance



- Sano – Centre for Personalised Computational Medicine – has been founded as a Centre of Excellence in development and deployment of decision support systems in medical science
- The Centre is based in Kraków, Poland, and operates as an independent foundation
- The Centre's activities involve integrating and processing large volumes of data related to medical research and decision support

https://sano.science

Sano - technical infrastructure

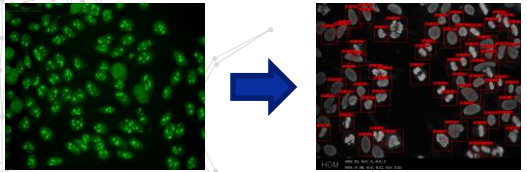
The Centre will support the work of **six independent research teams**, *each headed by a Research Team Leader and focusing on a specific area of computational medicine, as defined in Sano's International Research Agenda*

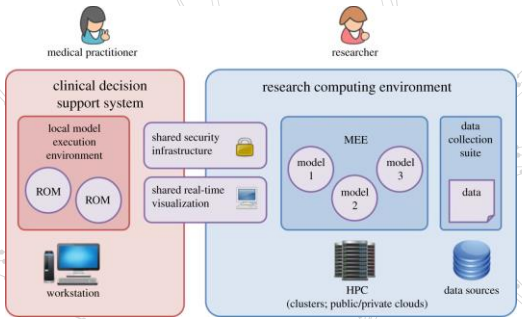
Sano provides general hardware and software infrastructure in support of collaborative R&D work:

- **Compute time:** *HPC grants: currently 300k CPU-hrs + 100k GPU-hrs*
- **R&D support staff:** *currently 6 part-time employees: ICT service team/Scientific Programmers team*
- **Software for code development:** *workflows, versioning, code audits*
- **Website:** *hosting and backup in Poland*
- **Basic hard- and software:** *notebooks, e-mail, cloud storage*

Initial Research Projects at Sano

- *Funded by IRAP Plus (Foundation for Polish Science)*
- *Goal: building research capacity, establishing collaborations via Sano*
- *Examples of collaborative projects initiated:*
  - *Image classification for immunological diseases (collaboration with CMUJ)*
  - *Modelling and simulation of Pulmonary Hypertension in (collaboration with USFD)*
  - *Feature selection for multiclass classification (collaboration with OMICRON UJ and Biobank Lab Univ. of Lodz)*
  - *Statistical methods of Machine Learning for Vasculitis (collaboration with CMUJ)*





Tools and Services

As part of its mission, Sano provides researchers with access to a comprehensive computational toolkit, including documentation and support staff

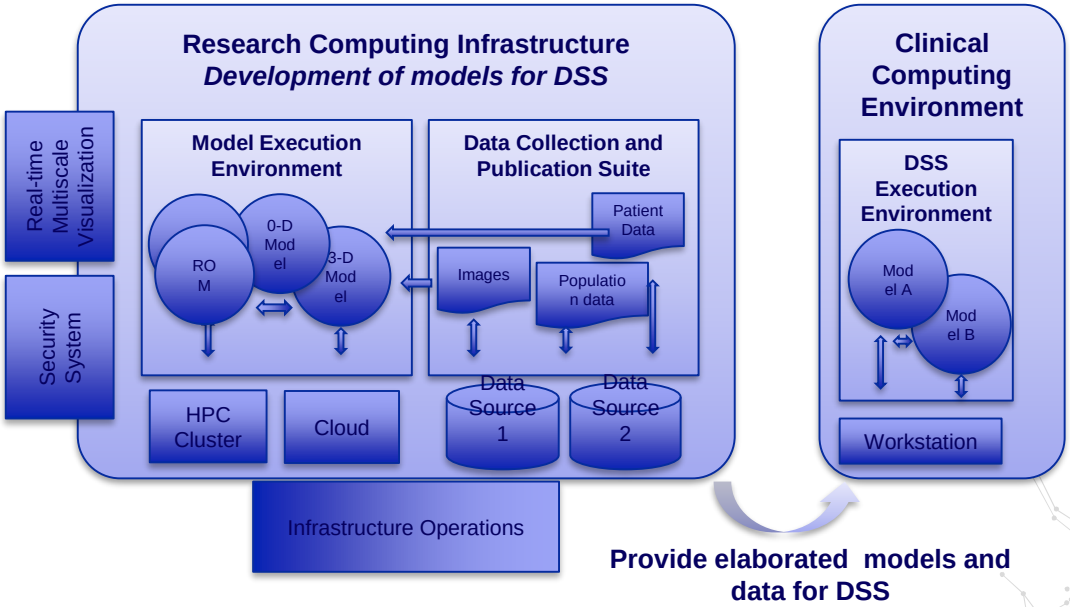
- **PLGrid access:** *user registration and certificate issuing in place – enables free-of-charge use of PLGrid resources, including a variety of proprietary (licensed) software, such as CFD solvers, Matlab, other specialized packages*
- **Model Execution Environment (MEE):**
  - *Comprehensive platform for deployment and reuse of computational models*
  - *Enables scientific computational pipelines running on HPC devices*
- *PLGrid also provides* **Virtual Machines** *for individual research tasks*

Challenge: how to integrate all of the available infrastructure into a coherent platform upon which large-scale (and medium-scale ;) scientific simulations can be executed in a secure, repeatable fashion?

Data access challenges at Sano

- Ensuring that data can be brought to where computations are performed – requires mechanisms for moving large data packets around the distributed environment
- Anonymization and secure storage of medical data sets (with GDPR compliance)
- „Right to be forgotten” – how to ensure that data is erased when the Centre no longer has permission to process it
- Easy upload and retrieval of input data and results to/from HPC platforms (which often do not provide convenient GUIs for end users)

Model Execution Environment – a computation and data integration platform

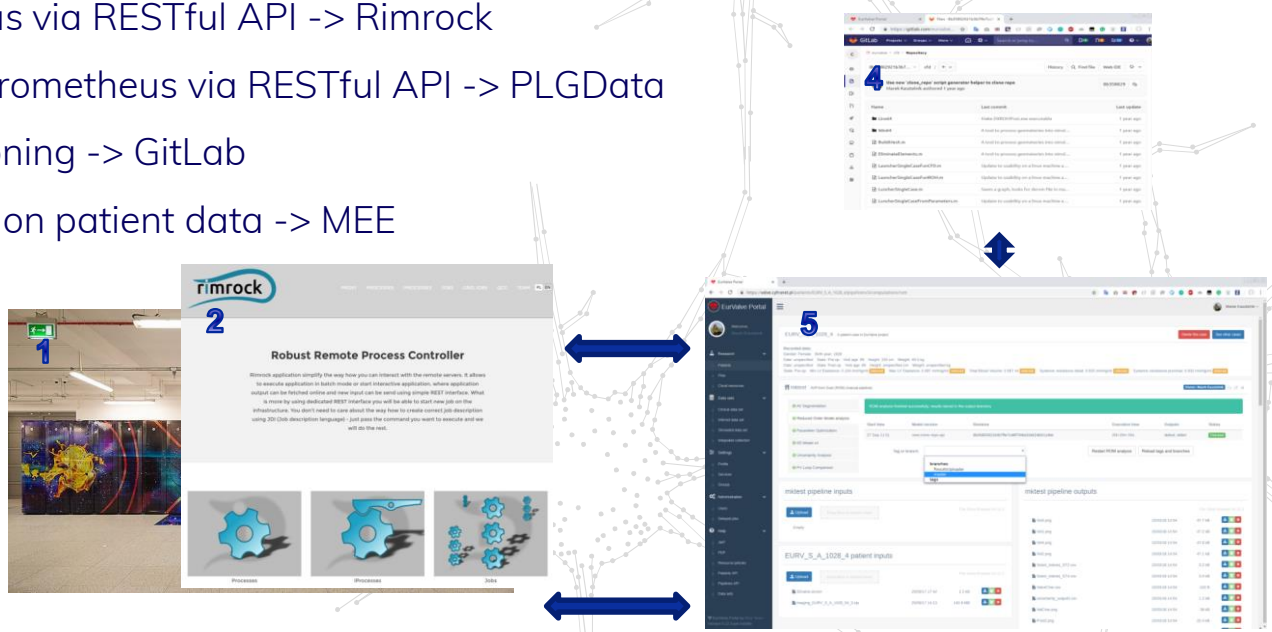


- The genesis of MEE dates back to the EurValve project, which called for extensive processing of large amounts of experimental data in order to prepare a data set which would later enable the operation of a clinical decision support system (DSS)
- The data concerned patients with valvular heart disease, and the goal of the DSS was to supply recommendations regarding their treatment
- Due to the fact that the DSS could not directly rely on HPC capabilities, heart models had to be precomputed in advance, by domain scientists (cardiologists)
- Thus, an environment was needed which would facilitate this task without requiring in-depth IT knowledge
- In addition to enabling execution of computational model on HPC resources, the tool called for integration of various data storage resources

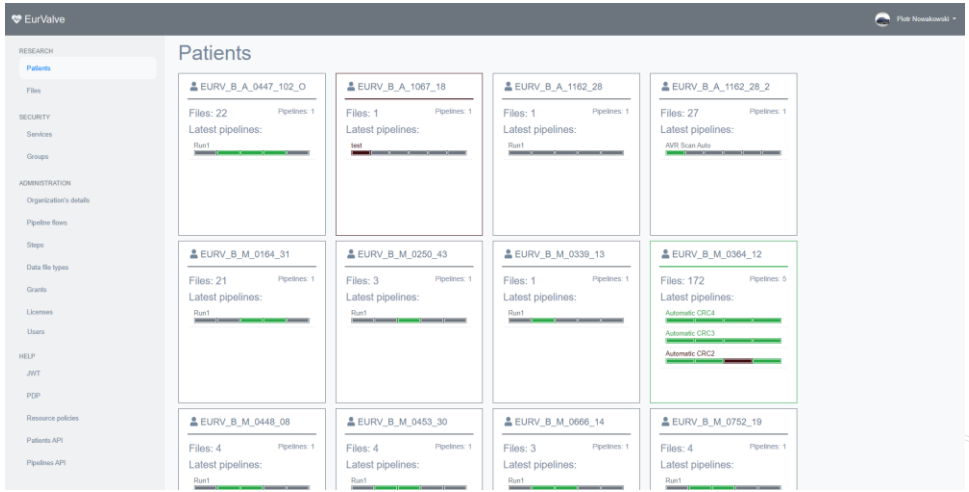
Processing scientific simulations – a conceptual framework

Five main elements required when processing patient cases:

1. HPC infrastructure -> Prometheus cluster
2. Running jobs on Prometheus via RESTful API -> Rimrock
3. Managing data stored on Prometheus via RESTful API -> PLGData
4. Model repository and versioning -> GitLab
5. Managing model execution on patient data -> MEE



Running simulations on integrated patient records



Each computational task is executed in the context of a specific case (e.g. a specific patient) and assumes the form of a pipeline

- A pipeline consists of one or more computational steps, each of which can be assigned required input and output files
- Steps are executed when their required input files are present in the HPC infrastructure – this enables sequential execution of steps; however, many steps may also be executed at the same time if they are capable of running in parallel.
- Each step corresponds to an executable which is stored and versioned in a GitLab repository – this enables repeatability and provenance tracking of output data
- Dedicated step templates are provided for stage-in and stage-out of input and output data for models which entail Big Data analyses.
- The environment provides graphical visualization of the execution status, and also enables platform users to define custom steps and custom pipelines.

Security

Sano uses mainly external infrastructure:

- *Central authentication: e.g. for communication and NextCloud storage*
- *Code and developer communication in private GitHub repositories*

Several security related processes have been implemented:

- *Several data security related SOPs in place*
- *Consulting agreement with an external data protection expert and appointment of a data protection officer (DPO)*

Involvement in Standards

Sano (or third party manufacturers) need to comply with EU **(CE marking)** *and other regulations to market clinical decision software systems (CDSS).*

- **Certification:** *MDR (Medical Device Regulation) 2017/745:*
  - *Applicable ISO standards: ISO 14971 (risk management), IEC 62304 (software lifecycle) and IEC 62366 (usability) in medical devices*
  - *Quality management system: ISO 13485, certified by a notified body, e.g. PCBC (Poland)*
- **Registration:** *National registries, e.g. URPL (Poland), EUDAMED database*
- *Additional regulations for international markets, e.g. HIPAA (US)*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533
and from the International Research Agendas Programme of the Foundation for Polish Science No MAB PLUS/2019/13.