The PIONEER Big Data Platform for Prostate Cancer

A unique infrastructure in Europe that enables a collaborative research environment to improve prostate cancer diagnosis, treatment, and care for patients and their families

PIONEER (Prostate Cancer DIagnOsis and TreatmeNt Enhancement through the Power of Big Data in EuRope) is a European Network of Excellence for Big Data in Prostate Cancer. Through unlocking the potential of Big Data and the generation of Real-World Evidence, PIONEER aims to change the prostate cancer landscape to improve the health and social care received by all prostate cancer patients and their families.

In this interview, the six experts Bertrand de Meulder, Daniel Kotik, Azadeh Tafreshiha, Peter Prinsen, Soundarya Palanisamy, and Mark Lambrecht present what the PIONEER Big Data Platform exactly is, how it is build up, how it works and in what way it contributes to answer the most important prostate cancer questions.

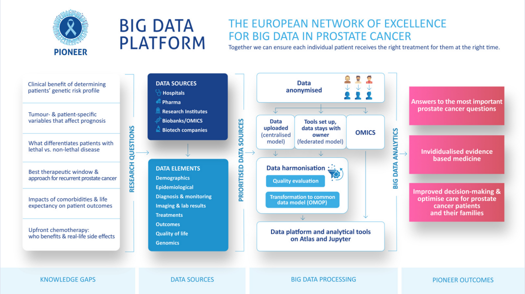
First of all, please briefly introduce yourselves and your role in the development, implementation, maintenance, and/or analysis of the platform

Dr Bertrand De Meulder: *"I am the science director at the European Institute for Systems Biology and Medicine (EISBM) in Lyon, France. In PIONEER, I am the academic leader of the Work Package 5 (WP5, Data Analytics), and a member of the WP4 (Data Platform). I have been involved in the design and implementation of the platform, the development of analytics, and the management of users."*

Daniel Kotik: "*I am a Research Software Scientist with a background in theoretical physics. We at the Center for Advanced Systems Understanding at the Helmholtz institute in Germany (CASUS/HZDR) provide the platform infrastructure and do the data on-boarding and user management. Our responsibility is to provide a reliable and sustainable platform which involves maintaining, securing and upgrading tasks on a regular basis."*

Dr Azadeh Tafreshiha: *"I am the project manager at the Real-World Data team at The Hyve. Under the PIONEER WP3 (Data Access and Sources) The Hyve has converted several datasets to OMOP CDM. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is an open community data standard, designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. Under WP4 The Hyve installed the central OHDSI-OMOP instance, which is now hosted at CASUS/HZDR."*

Click here to reach the PIONEER website


PIONEER
BIG DATA PLATFORM
THE EUROPEAN NETWORK OF EXCELLENCE FOR BIG DATA IN PROSTATE CANCER
Together we can ensure each individual patient receives the right treatment for them at the right time.
Clinical benefit of determining patients' genetic risk profile
Tumour- & patient-specific variables that affect prognosis
What differentiates patients with lethal vs. non-lethal disease
Best therapeutic window & approach for recurrent prostate cancer
Impacts of comorbidities & life expectancy on patient outcomes
Upfront chemotherapy: who benefits & real-life side effects
RESEARCH QUESTIONS
DATA SOURCES
Hospitals
Pharma
Research Institutes
Biobanks/OMICS
Biotech companies
DATA ELEMENTS
Demographics
Epidemiological
Diagnosis & monitoring
Imaging & lab results
Treatments
Outcomes
Quality of life
Genomics
PRIORITISED DATA SOURCES
Data anonymised
Data uploaded (centralised model)
Tools set up, data stays with owner (federated model)
OMICS
Data harmonisation
Quality evaluation
Transformation to common data model (OMOP)
Data platform and analytical tools on Atlas and Jupyter
BIG DATA ANALYTICS
Answers to the most important prostate cancer questions
Invididualised evidence based medicine
Improved decision-making & optimise care for prostate cancer patients and their families
KNOWLEDGE GAPS
DATA SOURCES
BIG DATA PROCESSING
PIONEER OUTCOMES

PIONEER Big Data Platform - 2

First of all, please briefly introduce yourselves and your role in the development, implementation, maintenance, and/or analysis of the platform

Dr Peter Prinsen: *“I am a clinical data scientist at the Netherlands Comprehensive Cancer Organisation (IKNL), the organization that maintains the Netherlands Cancer Registry (NCR). IKNL is an associate partner in PIONEER, which means that the NCR data can be used for studies in PIONEER. We share our data via the federated setup that means that we do not transfer our data to the PIONEER Big Data Platform but run the PIONEER studies ourselves and only share the results. We also support PIONEER studies by running early versions of study packages and providing the results for debugging and study protocol improvement.”*

Soundarya Palanisamy: *“I am an Industry Advisor with a focus on regulatory data science, based in Germany and am part of the Health and Life Sciences Practice at SAS. My role is to support in bridging the gap between industry and technology by implementing latest advanced analytics capabilities in the central PIONEER Big Data Platform for Prostate Cancer.”*

Dr Mark Lambrecht: *“I am a Director of Health and Life Sciences at SAS and have joined SAS in 2005. I lead a senior team focused on the health care and life sciences (HLS) ecosystem, and am part of the global health care and life sciences leadership at SAS. My interests include the impact of observational patient data, fostering data standards and enabling the industry to collaborate and reuse patient data to find new cures and to benefit all of humanity.”*

Can you briefly summarise what the PIONEER Big Data Platform is?

Dr de Meulder from EISBM:

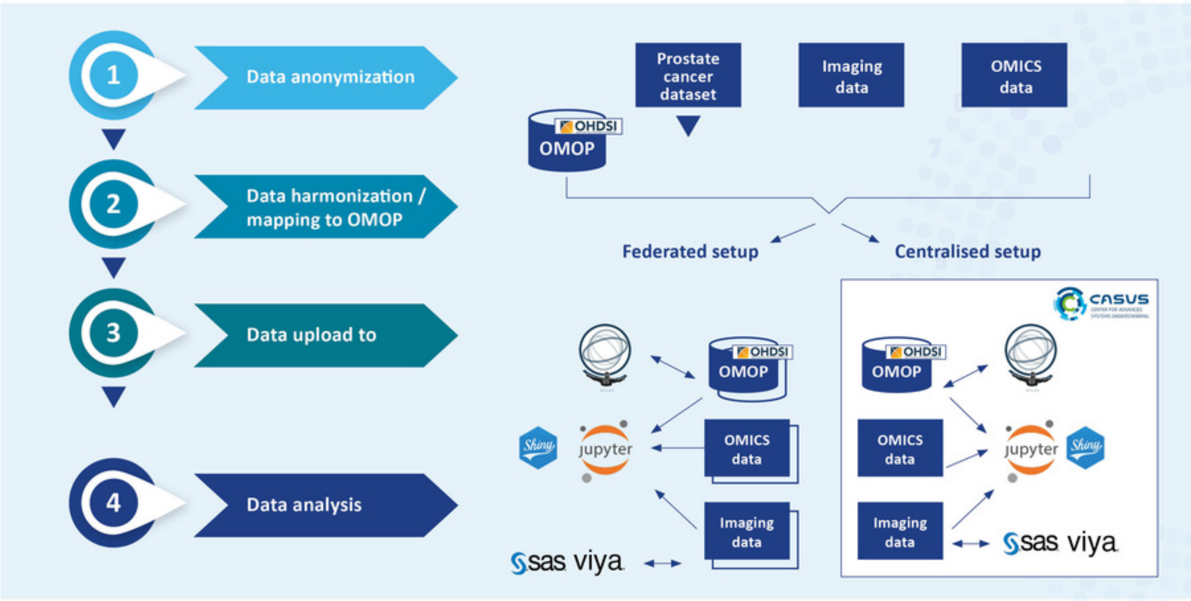
***"The PIONEER Big Data Platform*:**

- *Is a collection of Prostate Cancer (PCa) datasets (standardized clinical data, imaging and OMICS data), either hosted in the central sever or available through a federated data network as well as tools to explore the datasets hosted centrally in our CASUS/HZDR server (OHDSI ATLAS);*
- *Incorporates development tools for analytics – Github, JupyterHub, SAS;*
- *Offers a server for shiny apps to collate and present analytics results.”*

PIONEER Big Data Platform - 3

What does the data flow of Prostate Cancer data sets to the different platforms look like?

The PIONEER Big Data Platform offers a central and federated state-of-the-art Big Data analytic platform for prostate cancer. The platforms anonymous data is harmonised to the European Common Data Model, OMOP CDM, following all legal and ethical requirements and then analysed (see figure below).



“Central hosting model means the data owners share a copy of their dataset mapped to OMOP, that is physically hosted in the PIONEER central servers at HZDR, Dresden, Germany. Data owners are only responsible for helping in the mapping of their data and giving their consent for their data to be used in analyses.
Federated hosting model means the data owners keep the data mapped to OMOP in their own servers. The federated partner is responsible for mapping of their data, running analyses on their own computing infrastructure, sending results to be collated with the results of all other partners” **says Dr de Meulder from EISBM.**

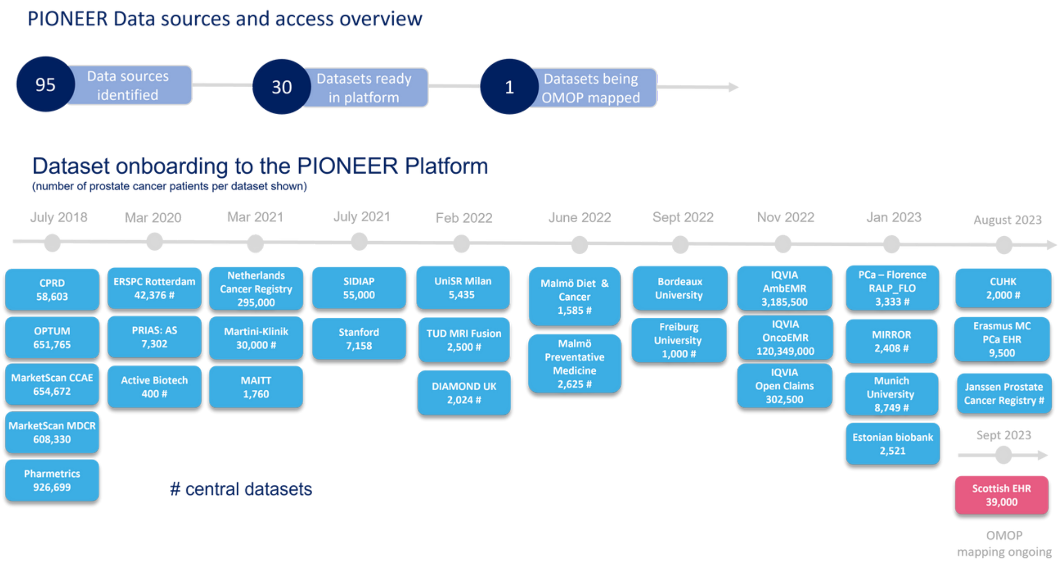
“SAS has deployed the latest version of its SAS Viya software in the PIONEER platform, expanding the analytic capabilities of the platform and supporting PIONEER researchers to accelerate clinical research into prostate cancer. SAS® Viya® is an extremely performant analytics engine that is designed to provide quick and accurate analytical insights in even the most complex environments with large amount of observational patient data. Data Scientists can leverage the power of open source and use various languages like SAS , R, Python – within this enterprise architecture”, **says Soundarya Palanisamy from SAS.**

"*The use of the OMOP CDM allows for a standardised approach to event tables and the use of curated vocabularies. In addition a new PIONEER specific concept set was created in the case the source terms did not correspond to any of the included vocabularies***", says Dr Tafreshiha from the Hyve.**

PIONEER Big Data Platform - 4

What kind of data are being deployed and where does the data come from?

The PIONEER Big Data platform brings together diverse datasets, contributed by our Partners and Associated Members, including the pharmaceutical industry, hospital electronic health records, large registry studies and interventional, and non-interventional trials. So far 30 data sets are available in the PIONEER Big Data Platform and one data sets are currently being mapped to the OMOP CDM.



PIONEER Data sources and access overview

95 Data sources identified → 30 Datasets ready in platform → 1 Datasets being OMOP mapped

Dataset onboarding to the PIONEER Platform
(number of prostate cancer patients per dataset shown)

- July 2018: CPRD 58,603; OPTUM 651,765; MarketScan CCAE 654,672; MarketScan MDCR 608,330; Pharmetrics 926,699
- Mar 2020: ERSPC Rotterdam 42,376 #; PRIAS: AS 7,302; Active Biotech 400 #
- Mar 2021: Netherlands Cancer Registry 295,000; Martini-Klinik 30,000 #; MAITT 1,760
- July 2021: SIDIAP 55,000; Stanford 7,158
- Feb 2022: UniSR Milan 5,435; TUD MRI Fusion 2,500 #; DIAMOND UK 2,024 #
- June 2022: Malmö Diet & Cancer 1,585 #; Malmö Preventative Medicine 2,625 #
- Sept 2022: Bordeaux University; Freiburg University 1,000 #
- Nov 2022: IQVIA AmbEMR 3,185,500; IQVIA OncoEMR 120,349,000; IQVIA Open Claims 302,500
- Jan 2023: PCa – Florence RALP_FLO 3,333 #; MIRROR 2,408 #; Munich University 8,749 #; Estonian biobank 2,521
- August 2023: CUHK 2,000 #; Erasmus MC PCa EHR 9,500; Janssen Prostate Cancer Registry #
- Sept 2023: Scottish EHR 39,000

central datasets

OMOP mapping ongoing

Please briefly describe the central data platform developed and hosted by the PIONEER project

Daniel Kotik: "*The PIONEER central data platform is a cloud-based, containerized, distributed system hosted on premise at HZDR's cluster infrastructure. The software stack builds on OHDSI tools (ATLAS/WebAPI), a JupyterHub with OHDSI customized R, Python and Julia kernels, a PostgreSQL database. It is complemented by SAS Viya4 running on Kubernetes***."**

Why is the PIONEER Big Data Platform unique?

Dr Bertrand De Meulder: *"The platform is unique as it is the largest (to our knowledge) collection of harmonized prostate cancer-related datasets, allowing access to records of millions of patients, together with state-ot-the-art analytics to exploit that resource."*

What is the aim of the paltform?

Dr Bertrand De Meulder: *"To be and remain the premier source of Real-World Evidence on prostate cancer, providing evidence of the highest quality possible to researchers, physicians and guidelines offices."*

What is the future of the paltform?

Dr Bertrand De Meulder: *"We want to ensure the sustainability of the platform for the foreseeable future and are gathering business ideas to fulfil that. We want to keep adding relevant data to our platform (e.g. linking European Health Record data updated frequently) and continue developing and performing analyses to extract knowledge from our data catalogue."*

PIONEER Big Data Platform - 5

In your opinion, what is the benefit of having such a Big Data Platform?

Soundarya Palanisamy

Big Data allows healthcare practitioners to drill down and learn more about their patients and the care they provide to them. Collecting high-quality data requires optimization of data collection tools in health care along with a reliable platform with extremely performant analytics engine to provide quick and accurate analytical insights in even the most complex environments – thereby help save lives, decrease costs, and improve the efficiency of operations.

Bertrand de Meulder

As a bioinformatician, I know that access to high-quality data is the most important key to success to answer research questions. Having an environment such as the PIONEER Big Data Platform makes it possible to answer important and relevant research questions that are otherwise impossible to tackle.

Daniel Kotik

It is often useful to collect dedicated data for a research project. One of the great aspects about PIONEER is that we want to show the potential of already existing patient data. In the field of cancer treatment, PIONEER proves that new and relevant scientific findings can also be achieved under the auspices of data economy and data protection.

PIONEER Big Data Platform - 6

In your opinion, what is the benefit of having such a Big Data Platform?

Peter Prinsen

The Big Data Platform makes (inter)national studies with other data partners much easier, since we do not have to harmonize data for every single study, and much safer, since we do not have to share patient-level data.

Mark Lambrecht

Our commitment at SAS is to complement and strengthen PIONEER's strategy to ready a digital environment where one is able to test detailed research hypotheses in an industry-grade analytics environment which will increase the speed of testing the new clinical trial data and foster collaboration between researchers.

One of the major challenges is that no single clinician has a view on the entire patient population and the impact of changing clinical guidelines on their individual patients. Having a platform to continuously enrich such kind of information will help identify correlations and signals that lead to a more precise approach for each patient. Patient-generated data, from a clinical perspective, improves outcomes by creating a more complete picture of the patient by including clinical, behavioral and biomarker data. Having statisticians and data analysts work in an environment where results can be shared in a "no-code" environment with clinicians is huge – but it takes everyone within PIONEER and beyond to overcome challenges on interoperability, semantics, ontologies, respect for patient privacy to release clinical outcomes using the best of what cloud-based environments have to offer today.