ESSAY

Introducing PIONEER: a project to harness big data in prostate cancer research

Muhammad Imran Omar[a], Monique J. Roobol[b], Maria J. Ribal[c], Thomas Abbott[d], Paul-Michael Agapow[e], Sonia Araujo[f], Alex Asiimwe[g], Charles Auffray[h], Irina Balaur[h], Katharina Bayer[i], Chiara Bernini[j], Anders Bjartell[k], Alberto Briganti[l], John-Edward Butler-Ransohoff[g], Riccardo Campi[c], Marinel Cavelaars[m], Bertrand De Meulder[h], Zsuzsanna Devecseri[n], Marc Dietrich Voss[n], Konstantinos Dimitropoulos[c], Susan Evans-Axelsson[k], Billy Franks[d], Louise Fullwood[o], Denis Horgan[j], Emma Jane Smith[c], Amit Kiran[d], Kati Kivinummi[p], Mark Lambrecht[q], Doron Lancet[r], Peter Lindgren[s], Sara MacLennan[a], Steven MacLennan[a], Maria Manuela Nogueira[h], Fredrik Moen[s], Maxim Moinat[m], Kishore Papineni[d], Christian Reich[f], Kristin Reiche[t], Stijn Rogiers[q], Claudio Sartini[g], Kees van Bochove[m], Femke van Diggelen[u], Mieke Van Hemelrijck[i], Hein Van Poppel[c], Jihong Zong[g], James N'Dow[c] and The PIONEER Consortium*

Affiliations:

a: Academic Urology Unit, University of Aberdeen, Aberdeen, United Kingdom
b: Erasmus MC, Rotterdam, The Netherlands
c: Guidelines Office, European Association of Urology, Arnhem, The Netherlands
d: Astellas, The Netherlands
e: Imperial College London, London, United Kingdom
f: IQVIA
g: Bayer AG, Berlin, Germany
h: Association EISBM, France
i: Translational Oncology and Urology Research, King's College London, London, United Kingdom
j: European Alliance for Personalised Medicine (EAPM), Belgium
k: Department of Translational Medicine, Lund University, Lund, Sweden
l: Department of Urology and Division of Experimental Oncology, Urological Research Institute, Vita-Salute San Raffaele University, IRCCS San Raffaele Scientific Institute, Milan, Italy
m: The Hyve, Utrecht, The Netherlands
n: Sanofi, France
o: Pinsent Masons, United Kingdom
p: Tampere University (TAU), Tampere, Finland
q: SAS, Belgium
r: Weizmann Institute, Rehovot, Israel
s: The Swedish Institute for Health Economics (IHE), Stockholm, Sweden
t: Fraunhofer IZI, Leipzig, Germany
u: ttopstart, Utrecht, The Netherlands

***The PIONEER Consortium include listed authors and the following collaborators:** Emelie Andersson, Heidi Arala, Anssi Auvinen, Chris Bangma, Danny Burke, Antonella Cardone, Joaquin Casariego, Guido Cuperus, Saeed Dabestani, Francesco Esperto, Nicola Fossati, Adam Fridhammar, Giorgio Gandaglia, Delila Gasi Tandefelt, Friedemann Horn, Johannes Huber, Jonas Hugosson, Henkjan Huisman, Michelle Jones, Andreas Josefsson, Olavi Kilkku, Markus Kreuz, Michael Lardas, Joe Lawson, Florence Lefresne, Stephane Lejeune, Elaine Longden-Chapman, Gordon McVie, Lisa Moris, Teemu Murtola, Charlie Nicholls, Karl H. Pang, Katie Pascoe, Marta Picozzi, Karin Plass, Pasi Pohjanjousi, Matthew Reaney, Sebastiaan Remmers, Paul Robinson, Jack Schalken, Max Schravendeel, Thomas Seisen, Angela Servan, Kirill Shiranov, Robert Snijder, Nesrine Taibi, Kirsi Talala, Derya Tilki, Thomas Van den Broeck, Zdravko Vassilev, Olli Voima, Eleni Vradi, Reg Waldeck, Ward Weistra, Peter-Paul Willemse, Manfred Wirth, Russ Wolfinger, Nazanin Zounemat Kermani

Abstract:

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, consisting of 32 private and public stakeholders from 9 countries across Europe. Launched by the Innovative Medicines Initiative 2 and part of the Big Data for Better Outcomes Programme (BD4BO), the overarching goal of PIONEER is to provide high-quality evidence on prostate cancer management by unlocking the potential of big data.

The project has identified critical evidence gaps in prostate cancer care, via a detailed prioritisation exercise including all key stakeholders. By standardising and integrating existing high quality and multidisciplinary data sources from prostate cancer patients across different stages of the disease, rich big data will be assembled into a single innovative data platform for research. Based on a unique set of methodologies, PIONEER aims at advancing the field of prostate cancer care with particular focus on improving prostate cancer-related outcomes, health system efficiency by streamlining patient management, and the quality of health and social care delivered to all prostate cancer patients and their families. The literature suggests there is underuse of effective treatments and overuse of ineffective treatment. For example, androgen deprivation therapy is sometimes overused in situations where it is not recommended. It is therefore crucial to identify the best treatment option for the individual patient.

Introduction

Prostate cancer is the second most common cancer in men by incidence in Europe, with 450,000 new cases diagnosed in 2018. Prostate cancer incidence varies five-fold, with the highest incidence in Northern and Western Europe, and the lowest in Central and Eastern Europe. The estimated incidence is highest in Ireland (189.3 per 100,000), whereas Albania (37 per 100,000) and Romania (47.2 per 100,000) have the lowest incidence (1). In 2018, the estimated numbers of death of prostate cancer were 107,300 for Europe (40 European countries), and 81,500 for 28 members countries of the European Union (1). Total annual estimated costs for treatment of prostate cancer in the first year following diagnosis is approximately €117 million in the UK. The figure is two- to three-fold higher in France and Germany (2). This economic burden associated with prostate cancer is predicted to dramatically increase in the coming years due to aging of the population, as around 85% of all cases of prostate cancer are diagnosed in men over the age of 65 years (1, 3, 4). Despite these numbers, up to now the level of funding for research is relatively low. For example, in 2018/2019, Cancer Research UK spend £13 million on prostate cancer research out of their total annual budget of £442 million (5). Therefore, progress made in prostate cancer research is limited when compared to other major cancer types. (1) For example, mortality statistics of Cancer Research UK indicate the mortality rate of breast cancer has been steadily declining, while the prostate cancer mortality rate is still on the rise (5).

Currently, several critical questions remain unresolved regarding the screening, diagnosis and treatment of prostate cancer patients, relating to various observations in prostate cancer epidemiology. First, prostate cancer incidence is variable across different European countries (37 to 189 per 100,000) (1). The differences in incidence rates of different racial and ethnic background confirms the involvement of genetic factors. However, environmental factors may also be implicated as the differences are also observed among men of the same genetic heritage who live in different European countries. Furthermore, inequalities in prostate cancer survival are also observed across the European Union. Estonia and Latvia have the highest mortality rates (37.3 per 100,000 and 35.7 per 100,000 respectively), whereas the mortality rates are the lowest in Spain and Italy (13.2 per 100,000 and 10.7 per 100,000 respectively) (1).

A variety of risk factors have been scrutinized for prostate cancer, including metabolic syndrome, obesity, dietary and genetics (6). However, the evidence on risk factors for prostate cancer remains inconclusive and, importantly, knowledge is lacking regarding patient characteristics (including molecular characterization) for optimal stratification of patients at time of diagnosis (6). Several diagnostic and prognostic tests for prostate cancer based upon molecular biomarkers have emerged, leading to a real challenge how to assess and prioritise these biomarkers (7). . Moreover, the variable pattern of prostate cancer screening and Prostate-specific antigen (PSA) testing across countries hinders a meaningful interpretation of available epidemiologic studies on the main risk factors for prostate cancer. Lithuania is among the few countries in the world where there is a national prostate cancer screening programme since 2006 (8). However, prostate cancer screening is considered one of the most controversial topics in urology, as there are different thresholds for screening frequency and intervals, and PSA thresholds for biopsy (9). This lack of knowledge means that safe identification of the candidates for active surveillance is suboptimal and similarly, predicting which patients will respond better to specific treatments remains difficult (6, 10).

Meaningful engagement of all key stakeholders is lacking in the processes that define the most important prostate cancer research questions that urgently need answers. The key stakeholders include clinicians, pharmaceutical companies, payers, and most importantly patients (11, 12). Ultimately, this negatively impacts research findings as the current focus in prostate cancer management may not be reflective of all different stakeholders.

Furthermore, knowledge gained in clinical practice (including knowledge informed by real life data) is not effectively implemented, with variability within and across European countries. PIONEER will collect data from different prospective and retrospective cohorts; patient registries; electronic health records; clinically recorded imaging data; patient encounters; problem lists; medication lists and histories; cancer therapy data; pathology reports, and; health-related quality of life outcomes. Ineffective implementation of knowledge gained in clinical practice, may lead to inequality in prostate cancer care, increased risk of short-term and long-term harms of interventions recommended to patients, as well as excess costs related to inappropriate management. A recent systematic review has identified geographical inequalities in the management of prostate cancer, and has highlighted that a better understanding of the complex social, environmental, and behavioural reasons for these variations is required (13).

PIONEER’s vision

The vision of PIONEER is to transform the management and clinical practice of prostate cancer across all disease stages (Stage I to IV) towards a data-driven and outcome-driven, value-based, and patient-centric health-care system. By applying advanced big data analytics, and developing a data platform of unparalleled scale, quality and diversity, PIONEER will empower meaningful improvement in clinical practice, prostate cancer disease-related outcomes, and health economic outcomes across the European health care landscape. PIONEER aims to bring together data from various sources including clinical, epidemiology, genetics, and health economics data. PIONEER will assemble, standardise, harmonise and analyse data from diverse populations of prostate cancer patients across different stages of the disease to provide evidence-based data for improving decision-making by key stakeholders (12). PIONEER brings together world-leading experts in clinical research, epidemiology, genetics, urology, big data science, health-economic research, private partners (EFPIA), and health-technology assessment.

Objectives of PIONEER

PIONEER has developed 8 individual work packages (WPs): project management and coordination (WP 1), 4 core research themes (WP 2-5) and 3 cross-cutting support themes (WP 6-8) (**Box 1**).

Approach and methodology

PIONEER will leverage existing valuable clinical prostate cancer datasets by bringing together a complementary group of world-leading clinical, epidemiology, genetics, urology, big data science, health economics, and health technology assessment (HTA) research experts, together with patient organisations, such as UCAN, Europa Uomo, and European Alliance for Personalised Medicine (EAPM) (12). The academic part of PIONEER is coordinated by the European Association of Urology (EAU), and their Guidelines Office, with financial support from the European Commission through the Innovative Medicines Initiative 2 (IMI2) (14), complemented by contributions from pharmaceutical industries and private partners of the European Federation of Pharmaceutical Industry Associations (EFPIA). In addition, the PIONEER consortium will build upon previous successful IMI projects and the other components of the BD4BO IMI2 framework (15). (**Figure 1**)

PIONEER has developed a dual approach, in order to use prostate cancer big data to develop an outcome-driven, value-based, and patient-centric healthcare system. First, PIONEER will identify critical evidence gaps in prostate cancer by combining the knowledge of academic and industry professionals and patients, thus enabling to focus the PIONEER working plan on a consensus list of research priorities and questions. Then, PIONEER will integrate, analyse, standardise and harmonise existing data from high quality and multidisciplinary data sources from prostate cancer patients across different stages of the disease into a single data platform (15, 16). To achieve this, PIONEER will use readily available, successful workbench and tools, such as tranSMART, OHDSI and the SAS open Platform, based on suitable data harmonisation techniques (OMOP Common Data Model) and advanced analytical methods. The advanced analytical methods may include machine learning, predictive modelling, multi-omics data integration methods, data visualisations as developed by The Hyve in the IMI1 funded project EMIF (the European Medical Information Framework) (17) and by the European Institute for Systems Biology and Medicine (EISBM (18)) and the Data Science Institute at Imperial College London (DSI-ICL (19)) in the IMI1 U-BIOPRED (20) and eTRIKS (21) projects.

Statistical analyses will be facilitated by utilising the KPMG (Klynveld Peat Marwick Goerdeler) Data Observatory within the DSI-ICL (19), thus enabling the analysis of complex datasets in a way that uncovers new insights in an immersive and multi-dimensional environment. To achieve this, the PIONEER statistical team will use the eTRIKS Analytical Environment (21), OHDSI R package open source (22) and SAS analytics software solutions (23).

Prioritisation of the most important questions in the field of prostate cancer

The EAU Prostate Cancer Guideline panel and other prostate cancer Key Opinion Leaders were contacted to identify the most important questions in the field of prostate cancer. Forty-four viable questions were identified. Afterwards, the PIONEER consortium performed a prioritisation survey among two stakeholder groups: healthcare professionals including pharmaceutical companies and prostate cancer patients.

In total, 73 healthcare professionals and 57 patients participated in round one of the surveys. The results were analysed by calculating the percentage of respondents scoring each question as not important, important or critically important.. Twelve additional questions were proposed during the first round. For the second round the patients' surveys were also translated into French, German, Italian and Spanish. 49 healthcare professionals and 169 patients (including 53 English; 19 French; 31 German; 53 Italian; 13 Spanish) participated in round two of the surveys. These 56 questions (44 questions from round one and 12 additional questions from round two) were then re-ordered according to the highest percentage for "critically important". The questions covered all stages of prostate cancer focusing on various aspects of the condition including screening, diagnosis, risk stratification including the genomic profile, treatment, and complications of treatment. The detailed results will be presented in a separate publication, but in meantime are being used to inform PIONEER consortium, so that the stakeholder groups' priorities are met in an accountable and transparent way.

WP1: Project management and administration

PIONEER WP1 ensures the efficient management of the consortium, the progress of the project towards the planned objectives and deliverables. Implementation of an appropriate governance structure that allows efficient interaction of the different stakeholders, including management bodies as well as external scientific and ethical advisory boards, and preparation of decision-making by the management bodies are crucial aspects of the consortium management. Given the large number of academic organizations, institutes and private companies participating in PIONEER (n=32), a major portion of coordination work will be required to ensure an appropriate flow of information between the different WPs, to facilitate internal communication between the participants and to coordinate external stakeholder interactions supporting dissemination and communication **(elaborated below in WP7:**

Dissemination and communication). Furthermore, the linkage of PIONEER with other programmes of the BD4BO initiative and sustainability of the project's outcomes beyond the project duration are integral objectives.

WP2: Disease understanding and outcome definition

The aim of WP2 is to develop standardised definitions and measurements of prostate cancer outcomes and diagnostic, predictive, prognostic, and therapeutic factors (DPPTs) across the different stages of prostate cancer care, and to consider the opinions of key stakeholders in this process (**Box 1: PIONEER Research Objectives)**.

To date, many prostate cancer outcomes and DPPTs have been arbitrarily defined and, in the case of DPPTs, have mainly been investigated in single cohorts. Even in Randomized controlled trial (RCT) data, heterogeneity of outcome definition and measurement limits critical appraisal and statistical synthesis of data across sources. This means that analyses cannot harness the power and precision of all available data. Healthcare providers must choose from a wide array of diagnostic tools and treatment modalities but due the lack of consensus on the most important prostate cancer-related outcomes and DPPTs, clinical practice decision-making is more dauntingly complex than it should be. This contributes to unacceptable inequalities for prostate cancer patients observed throughout Europe. Therefore, confirmation of the effectiveness of treatments, or the accuracy of diagnostic tests, or the utility of predictive biomarkers, can be known with confidence only if the prostate cancer outcomes and DPPTs become standardised. These standardised definitions will be thus applied to the large studies contributing data to the PIONEER platform (including data from patients with different lifestyles and from a range of healthcare systems), in order to identify outcomes that will allow to discern which patient will benefit most from what treatments, and to facilitate both drug development and more appropriate patient care.

The objectives of WP2 are to reach a consensus for each stage of prostate cancer on which outcomes are the most important for stakeholder groups including healthcare professionals and patients, how they should be defined and measured, what DPPTs are the most important for various stakeholders, and how they should be defined and measured.

First, for the outcomes standardisation work we will update and integrate existing Core Outcome Set (COS) developed using the COMET and ICHOM processes (24-26). We will involve both groups in this task and create an up-to-date COS for use within PIONEER and for future effectiveness trials and clinical audit. We will also survey which DPPTs already exist for the different stages of prostate cancer care (i.e. screening, diagnostic, staging and treatment activities) and assess which ones have discriminatory and predictive value. For all reviews we will follow the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (27). These systematic reviews will map current practice and complexities involved in diagnosis, prognosis and management of men with prostate cancer and overview the outcomes currently used in research.

In addition to assessing published literature in our systematic reviews, we will also evaluate the data collected in the different data resources of PIONEER. This process will result in a structured database of verbatim outcome names, definitions and measures. The outcomes database will be categorised according to the generic Williamson Clarke taxonomy (28), with additional prostate cancer specific outcomes and definitions provided **(elaborated below in WP4: Data platform)**. This will structure and homogenise the available COSs.

Second, the group will prioritise the identified outcomes and DPPTs for each stage based on the preferences of different stakeholders involved (i.e. patients and their family/partner/carer, HTAs, payers/insurance groups, pharmaceutical industry, etc.) using a modified Delphi consensus-building process as advocated by the COMET initiative (29), and demonstrated in other prostate cancer specific studies (24, 30).

The last step will be to identify how to measure the identified COS and DPFs. Currently, selecting an appropriate outcome measure instrument is challenging given the comprehensive list of outcome sets WP2 is developing. There is often no single best measurement defined for the different outcomes so the optimal definition for clinician reported outcomes (e.g. progression or recurrence) may need to be based on consensus. In addition, the optimum tools to be used for patient reported outcomes (e.g. urinary function, quality of life) may rest on the assessment of the tool's content validity within the target population, then on other psychometric properties, and the assessment of its feasibility in research and practice. Ultimately, WP2 aims to develop a pragmatic way to select the appropriate definitions and measurements.

The final definitions and measurements will be used as a) the basis of harmonisation of the outcomes definitions data within PIONEER datasets, and b) the COSs recommended to be collected as a minimum in future routine data collection, observational studies and clinical trials. The WP2 has already made substantial progress in standardising and harmonising outcomes for the interventions of patients with localized and locally advanced prostate cancer. The PIONEER WP2 first identified all reported outcomes (such as overall survival, prostate cancer specific survival) from clinical trials of interventions by conducting systematic reviews. This was followed by expert group consensus meetings with clinicians, patients, academics and industry representatives, where the identified outcomes from clinical trials were discussed in detail, to standardize terminology and to recommend core-outcomes set for localized prostate cancer, that can be used for future research including clinical trials and studies. The WP2 will develop core-outcomes sets for non-localized prostate cancer as well. WP2 is currently working on the systematic review protocol of diagnostic and prognostic factors for all stages of prostate cancer.

WP3: Data access and sources

WP3 aims to identify, approach and negotiate appropriate data access agreements with a variety of potential holders of high-quality, prostate cancer-based datasets across European (and non-European) patient populations (**Box 1: PIONEER Research Objectives)**. WP3 will collect, standardise and harmonise existing prospective and retrospective data into a single innovative data platform developed by WP4 **(elaborated below in WP4: Data platform)**. To effectively implement the WP3 workplan, subgroups were formed within WP3.

As part of the initial proposal for the PIONEER consortium, 27 potential data providers were identified. This number has since grown to over 60 data sources and is expected to continue to grow as new sources are identified. Potential data contributors include large clinical practices and medical centres, life sciences companies, data aggregators and payers/governments.

WP3 will contact biomedical institutes and hospitals holding clinical data, assess their willingness to participate by obtaining a signed letter of intent and collect information about the contents of their database(s) by filling in a data contributor Fact Sheet. These Fact Sheets form the basis for the 'clinical fingerprint' (omics data type relevant to prostate cancer) used in the EMIF central metadata catalogue developed by The Hyve (WP4) (17). .

Once the data providers' intent to participate is confirmed, WP3 will begin to negotiate appropriate Data Access Agreements (DAAs). The DAA templates are based on other IMI project agreements (*i.e.* HARMONY (31)) modified by Pinsent Mason Associates **(Elaborated below in WP8: Legal, ethical issues and governance)** to suit PIONEER data providers' needs. To encourage participation in the PIONEER consortium, the DAAs outline the policies and procedures under which their data can be accessed and analysed. The DAAs will also include sections which satisfy country-specific General Data Protection Regulations (GDPRs), data governance and value propositions tailored to each type of data provider. Suggested value propositions include authorship, benchmarking, clinical decision-making, transparency initiatives, technical support and networking opportunities. In exchange for signing the agreements, data providers are given certain rights and privileges (*e.g.* the right to propose research questions, request authorship and opting out of study participation) along with accepting certain obligations (*e.g.* a commitment to participate in studies whenever possible).

Upon signing the DAAs, WP3 will work to convert, harmonise and map the data sets into a common data model similar to other IMI projects, e.g. EMIF, using a variety of approaches and software while also maintaining security and consistency. The multiple data sets will be linked to form the PIONEER platform used for subsequent analyses.

The overall objective of PIONEER is to establish a long-term sustainable research network, with established policies and procedures for the access and analyses of big data from multiple sources. WP3 is establishing data management plans to support this sustainability goal that include options for data providers to continue their participation after the initial funding phase or withdraw their participation and have their data appropriately decommissioned from the PIONEER platform.

The biggest challenge is centred around the development of an appropriate data access framework which will motivate contributors to participate, satisfy GDRP and privacy regulations while allowing meaningful research collaborations.

WP4: Data platform

PIONEER WP4 will develop a pan-European data-sharing platform and adopt a two-pronged approach to address the project needs: a) a platform that can access population-based registry data such as electronic health records, and b) a platform that can handle rich clinical and omics data for translational analysis by WP5 **(elaborated below in WP5: Data analytics).** To achieve this, the project will build upon and use approaches developed in a number of other IMI projects, such as IMI1 EMIF (17) and IMI1 eTRIKS (21).

For data integration and analysis of longitudinal prostate cancer registries, PIONEER will use the OMOP and OHDSI (22) technology, while for cohort studies that also include omics data besides deep phenotypic and clinical data, the tranSMART (32) technology will be used (**Figure 2**). Components from both technologies are still under development.

WP4 will also use the EMIF catalogue to facilitate centralised storage, management and sharing of metadata of available prostate cancer data sets. It will provide a list of all the data sources registered by PIONEER, through a portal and search tools, to enable potential data users to discover data sources that are most relevant to their research needs, according to a variety of data source and dataset descriptors, and to support the access request process.

All data will be harmonised (tranSMART) or standardised (OHDSI-OMOP) before being loaded in the platform of choice (**Box 1: PIONEER Research Objectives)**. The open-source/available tools of the IMI1 eTRIKS and IMI1 EMIF projects will be used for harmonisation of data that are loaded in tranSMART.

We have envisioned distinct possibilities depending on the nature of the data (centralised vs. decentralised/federated). In particular, central installation of tranSMART and OMOP-ATLAS will be chosen for data that may leave the source server or repository, while federated installations of OMOP-ATLAS will be chosen for data that may not leave the data provider's premises.

We envision that certain federated data sources contain omics information. Currently, there is no existing platform that support federation of omics data and *de novo* development would be beyond the resources and time available. If it becomes a clear need in PIONEER, we have the choice to either adapt tranSMART to support federated analysis or support omics in the federated OHDSI platform. Within OHDSI, there is a workgroup aimed at creating support for analysis of genomics data.

WP5: Data analytics

PIONEER WP5 is in charge of planning, performing and evaluating the bioinformatics and systems biology analyses to answer PIONEER research questions.

The team in WP5 will provide a unique toolkit of standard and cutting-edge analytical methods for the analysis of big data, both from open-source and industry-developed methods (**Box 1: PIONEER Research Objectives)**. Research questions and core outcome sets have been identified in PIONEER's survey conducted by WP2. Each of the research questions that PIONEER will tackle will require different tools and analytic workflows that will be provided by WP5, through the centralised tranSMART omics platform built by WP4 (**Figure 3**).

Data analytic workflows in WP5 are built around two main sources: open-source software (mainly R packages) and commercial software (SAS) (23). Each source has its advantages and limitations with regards to technical possibilities, user-friendliness, built-in visualisation capabilities, etc. We envision that the different research questions will require different types of analytical methods and that different sources will be better suited to meet those requirements. It is also our expectation that open-source and commercial analytical methods will feed each other to generate the best possible results for the benefit of the project and the patients.

PIONEER will achieve its aims by performing the following tasks. First, we will write, evaluate and circulate data analysis plans and standard operating procedures for data analysis. We will explore and characterize the demographic and geographic data available to us through the Data Observatory at the ICL and the visual capabilities of the data analysis platform. In this process, we will constantly focus attention on the lookout for data error and outliers to seek the cleanest and most reliable data possible. We will then perform initial analyses by generating data descriptive statistics to assess the existing predictive models in our dataset and decide on our benchmarks and internal validation schemes. This will allow us to use advanced analytical methods with confidence, including but not limited to, multiple omics data analysis (33), topology data analysis (34), regression modelling (e.g. OPLSDA method (35)), genetic risks prediction (36), random forest machine learning (37), etc. As the databases will be a collection from disparate populations and/or database sources, meta-analytic techniques will be employed to account for between- and within-population variability and heterogeneity (38, 39). Making sense of the results will be done with the help of knowledge bases including gProfiler (40), MalaCards (41) and STRING (42). The various predictive models will be combined into a predictive algorithm for the use of health specialists. We will demonstrate the improvement of our newly developed models on the benchmarks and related to the economic burden of prostate cancer management, and provide user-friendly scores and evaluating schemes for the physicians and patients benefit [nomograms (43), against over-diagnosis and over-treatment (44)]. Finally, we will provide recommendations and guidance documents, disseminated to professional and patient organisations (*e.g.* EAU, International Shared Decision-Making Group, Europa Uomo, Movember) in collaboration with PIONEER WP7 **(elaborated below in WP7: Dissemination and communication).**

The list of analytical tools that are expected to be of use in PIONEER is still being refined. However, there will be a heavy need for predictive modelling and machine learning (random forests, linear modelling, support-vector machines, partial least square regressions). Visualisation will be provided both at the level of the data platform (either in tranSMART or OHDSI) directly and with dedicated software included in the SAS suite and in R packages. We will also monitor and make use of developments in the broader computational systems biology community as they become available to use. High-performance computing, bringing big data analytics capabilities when needed to answer PIONEER research questions will be provided through a SPARK infrastructure hosted at the DSI-ICL. Finally, WP5 will make use of the developments, insights and experience of other research projects through partnerships, research seminars and the projects members' experience, either from IMI projects [eTRIKS, B4B (Brains for Brain), EMIF, parallel BD4BO IMI projects] or from the industry partners' internal knowledge and developments.

WP6: HTA regulator

Through WP6, the PIONEER project will seek to develop, and also validate, a framework for innovative technologies in prostate cancer using real-world evidence (**Box 1: PIONEER Research Objectives)**. The latter involves using various health data in real time to help healthcare professionals make better and quicker decisions. Real-World data has been defined as "an umbrella term for different types of healthcare data that are not collected in conventional randomized controlled trials... including patient data, data from clinicians, hospital data, data from payers and social data" (45).

Many HTA and payer groups think of real-world evidence as having much potential, but alignment is still necessary. PIONEER will work with such bodies as well as regulators to establish minimum evidence requirements while identifying, at an early stage, potential uncertainties requiring extra data. On top of this, PIONEER will seek to develop reference models for use in economic evaluations and, as a key objective, will explore whether it can develop a core set of reference models for different stages of prostate cancer, or an overarching modelling framework. This is necessary in order to explore the impact of new technologies at single points along the pathway, as well as looking at treatment sequences, as the disease progresses through its multiple stages.

Effective evaluation of the medical, social, economic and ethical issues of products in a systematic, transparent, unbiased, robust manner will promote safe, effective, health policies that are patient-focused and obtain best value - whether at the time of launch or their usage in real-life circumstances. Adapted tools and openness to evidence produced by methods other than classic RCTs will be helpful. In the case of adaptive clinical trials, real-world evidence is crucial. Medical Adaptive Pathways to Patients, known as MAPPs, have been tested in a European Medicines Agency (EMA) project, and will be used in PIONEER.

MAPPs are described as a prospectively planned process, starting with the early authorisation of a medicine in a restricted patient population, followed by iterative phases of evidence gathering and adaptations of the marketing authorisation to expand access to the medicine to broader patient populations. The keywords here are the 'iterative phases of evidence gathering', which should use real-world evidence to detect patient responses to new therapies in a real-time setting.

Meanwhile, many of those mobile applications that we are all now using are essentially gathering real-world evidence on a daily basis. The more advanced health applications can provide this while also running simultaneous comparative efficacy trials against existing therapies. These applications could also solve issues surrounding data interoperability, given that a data standard is already in place when using iOS or Android platforms. In theory, these real-time datasets could then be sent, for example, as standard XML files to any internet database in Europe and beyond.

At the time of writing, several national health databases have been providing an opportunity to search, identify, and target (pseudo) anonymised patient data. These data can become available to healthcare professionals offering them integrated real-time updates in the case of national health records.

The practical benefit could be that such databases may allow a radical reduction in the development time usually needed in RCTs. Through the stakeholder working group, PIONEER will propose policy recommendations to develop this area in a structured manner.

WP7: Dissemination and communication

Through WP7, PIONEER will communicate information to the public about the project and its implementation status by providing comprehensible, educational, and operable information on PIONEER's outcomes to all relevant stakeholders including policy-makers as they play a key role in shaping the research agenda, thus facilitating the implementation and adoption of PIONEER's results (**Box 1: PIONEER Research Objectives**).

WP7 will ensure effective communication within the consortium. Effective internal communication between consortium partners is of utmost importance. Each partner must be informed on the progress of the entire project and share common goals and objectives. WP7 will coordinate communication activities with other relevant research and stakeholder networks and provide for the dissemination of project developed platforms for use by the wider scientific community.

Optimal and effective dissemination of PIONEER results is essential for the ultimate success of the project. Our vision is that PIONEER outcomes will influence current (and future) research agendas, clinical development processes, and reshape current clinical practices based on up-to-date evidence derived from real-life data. To achieve these objectives, PIONEER will require the support of all relevant stakeholder groups and to accomplish this, an effective communication and dissemination strategy has been developed. This strategy forms the base of all PIONEER communications directions and will be periodically revised to reflect stakeholders' feedback relating to the different communication tools and channels. Primarily, PIONEER dissemination approach is two-fold with the initial phase focused on increasing general awareness of the project and the second phase geared towards tailored messages delivered to specific stakeholder audiences.

The PIONEER project website has been developed and was launched during the projects kick-off meeting (14th May 2018), under the registered domain https://prostate-pioneer.eu. In addition, a PIONEER Twitter account was created in May 2018 (@ProstatePioneer).

Identification of PIONEER target audience is a key step for successful dissemination of the project outcomes. Successful identification and engagement of all relevant stakeholders could be a potential challenge. Knowing our target audiences involves knowing the specific needs of the individual audience and not just the message we want to convey. To overcome communication barriers, it is important to determine the medium through which PIONEER will communicate with different target audiences and the timing of message delivery. Information needs to be of good quality, timely, contextually relevant and appropriate to the intended audience. Furthermore, the early involvement of all relevant stakeholders is a key enabler: being actively involved in the design and prioritisation of the research questions addressed throughout the project will help greatly in ensuring ongoing stakeholder engagement enabling PIONEER to meet its objectives.

WP8: Legal, ethical issues and governance

In PIONEER WP8, we will be seeking to: (a) map best practices and related issues concerning governance of big data solutions in healthcare, (b) consolidate learnings to assist the development of a sustainable governance structure covering issues that may arise from the use of big data collected from human participants (e.g. use of personal data, patient confidentiality, patient consent and data ownership), (c) facilitate responsible use of data by providing advice and guidance to assist all project participants to understand and hence be better able to comply with relevant legal, regulatory and ethical requirements on privacy and data protection; (d) coordinate the activities with other IMI2 BD4BO projects to share, test and evaluate ideas; and (e) provide guidance on dealing with informed consent forms in the event that the project includes prospective data gathering (**Box 1: PIONEER Research Objectives)**.

Currently, there are no universally accepted best practices for involvement of patients and public in such initiatives, and there are still some unanswered questions to be addressed. Following the existing debate on data protection and the role of patients in clinical research, PIONEER will establish an Ethical Advisory Board. WP8 will identify the existing best practices to involve not only co-participants and other stakeholders but also patients and their organisations in the definition and solution of relevant ethical and legal issues. WP8 will ensure respecting the privacy rights of the people whose personal data are processed, the clinical profession duty of confidentiality and the protection of the interests of participants and researchers (46).

The timing of the PIONEER project coincides with the implementation of the GDPR, which came into force in May 2018. This is the most significant change in data privacy law over the past 20 years and creates a challenge to the project in that it allows for individual member states to choose to apply or to derogate from certain aspects of the GDPR. One of these areas is the secondary use of healthcare data for research purposes and means that the project must understand and deal with differing compliance requirements in different member states. We will be exploring ways to address this, including avoiding the transfer of personal data by using a federated data model and/or by using anonymisation so as to take the relevant data outside of the GDPR framework. Thus enabling implementation of a mixed model in which part of the data, could be handled efficiently and securely in a centralised data and knowledge management platform (46).

Planned outcomes of PIONEER

PIONEER will assemble, standardise, harmonise and analyse high-quality big data from diverse populations of prostate cancer patients across different stages of the disease to provide evidence-based data for improving decision-making by key stakeholders. This will lead to meaningful improvement in clinical practice, prostate cancer disease-related outcomes, and health-economic outcomes across the European healthcare system. Some of the planned outcomes of PIONEER are listed below (Figure 3):

- Consensus on the most important prostate cancer outcomes (WP2: Disease understanding and outcome definition)
- Identification of critical evidence gaps in prostate cancer (as detailed above under: Prioritisation of the most important questions in the field of prostate cancer
- Standardisation of outcome definition and outcome measures outcomes
- New insights on improved stratification
- Improved standardized care pathways with known better predictable outcomes

Challenges in PIONEER

The PIONEER project may come across some important challenges. These challenges will not only be of legal and ethical nature, but we will come across some methodological challenges as well, such as data quality, data inconsistency, limitation of observational studies, and analytics issues. The use of big data in medical research and in healthcare systems raises complex ethical issues, which have significant implications for policy and legal frameworks. This includes challenges ranging from consent, data privacy, cyber security, to wider social aspects of the uses to which patient data may be subject. PIONEER has established an appropriate framework (noting the potential for different applications of certain regulations between different member states) to ensure that data access, release and linkage, and governance of combined datasets of the consortium are addressed in a manner compliant with legal, regulatory and ethical requirements, and that relevant WPs are handling data accordingly so that patient trust.

Future directions

By 2023 the PIONEER project will deliver essential lessons for targeted care and management of prostate cancer patients. It will house a central data hub supporting a network of interdisciplinary personnel, to address critical scientific questions.

The success of this journey depends on several key factors, including logistical aspects (public and private collaboration), data availability, access to data, data quality and harmonisation, as well as the adoption of a new generation technology into the platform.

The project will highlight the benefits and power of big data to answer important clinical questions. Transparency and strict legal oversight will guarantee for protection of patients' privacy. Our aspiration is to include data from as many countries as possible to represent the prostate cancer patient population worldwide.

The biggest challenges of PIONEER will likely be to maintain this work and platform accessible to researchers and clinicians looking for answers to better manage their difficult patient cases.

Inclusion of the most appropriate outcome measures as well as relevant economic aspects, can guide payers to make the right reimbursement decisions.

The potential of PIONEER is immense, with the key for success being a strong foundation. This unique collaborative structure and outstanding commitment from all participants will hopefully set a model for other similar big data projects for the benefits of patients, healthcare professionals, and other relevant stakeholder.

References:

1. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer. 2018;103:356-87.

2. Smith-Palmer J, Takizawa C, Valentine W. Literature review of the burden of prostate cancer in Germany, France, the United Kingdom and Canada. BMC Urol. 2019;19(1):19.

3. Patel AR, Klein EA. Risk factors for prostate cancer. Nat Clin Pract Urol. 2009;6(2):87-95.

4. Luengo-Fernandez R, Leal J, Gray A, Sullivan R. Economic burden of cancer across the European Union: a population-based cost analysis. Lancet Oncol. 2013;14(12):1165-74.

5. Cancer Research UK. Available from: https://www.cancerresearchuk.org [Accessed 5th March 2019]

6. Campi R, Brookman-May SD, Subiela Henriquez JD, Akdogan B, Brausi M, Klatte T, et al. Impact of Metabolic Diseases, Drugs, and Dietary Factors on Prostate Cancer Risk, Recurrence, and Survival: A Systematic Review by the European Association of Urology Section of Oncological Urology. Eur Urol Focus. 2018.

7. Kohaar I PG, Srivastava S. A Rich Array of Prostate Cancer Molecular Biomarkers: Opportunities and Challenges. Int J Mol Sci. 2019;20(8):1813.

8. Gondos A, Krilaviciute A, Smailyte G, Ulys A, Brenner H. Cancer surveillance using registry data: Results and recommendations for the Lithuanian national prostate cancer early detection programme. European Journal of Cancer. 2015;51(12):1630-7.

9. Ilic D, Djulbegovic M, Jung JH, Hwang EC, Zhou Q, Cleves A, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. BMJ. 2018;362:k3519.

10. Brookman-May SD, Campi R, Henriquez JDS, Klatte T, Langenhuijsen JF, Brausi M, et al. Latest Evidence on the Impact of Smoking, Sports, and Sexual Activity as Modifiable Lifestyle Risk Factors for Prostate Cancer Incidence, Recurrence, and Progression: A Systematic Review of the Literature by the European Association of Urology Section of Oncological Urology (ESOU). Eur Urol Focus. 2018.

11. Health Europa. Available from: https://www.healtheuropa.eu/enhance-prostate-cancer-care/85990/ [Accessed 24th July 2019].

12. PIONEER. Prostate Cancer DIagnOsis and TreatmeNt Enhancement through the Power of Big Data in EuRope. Available from: https://prostate-pioneer.eu [Accessed 5th March 2019].

13. Dasgupta P, Baade PD, Aitken JF, Ralph N, Chambers SK, Dunn J. Geographical Variations in Prostate Cancer Outcomes: A Systematic Review of International Evidence. Front Oncol. 2019;9:238.

14. IMI. Innovative Medicines Initiative (IMI). Available from: www.imi.europa.eu [Accessed 5th March 2019].

15. BD4BO. Big Data for Better Outcomes. Available from: http://bd4bo.eu [Accessed 5th March 2019].

16. Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, et al. Making sense of big data in health research: Towards an EU action plan. Genome Med. 2016;8(1):71.

17. EMIF. European Medical Information Framework. Available from: http://www.emif.eu [Accessed 5th March 2019].

18. EISBM. European Institute for Systems Biology and Medicine. Available from: http://www.eisbm.org/ [Accessed 5th March 2019].

19. London IC. Data Science Institue. Available from: https://www.imperial.ac.uk/data-science/about-the-institute/ [Accessed 5th March 2019].

20. U-BIOPRED. Unbiased BIOmarkers in PREDiction of respiratory disease outcomes. Available from: https://www.europeanlung.org/en/projects-and-research/projects/u-biopred/home [Accessed 5th March 2019].

21. eTRIKS. eTRIKS Harmonisation Services. Available from: https://www.etriks.org/etriks-harmonisation-services/ [Accessed 5th March 2019].

22. OHDSI. Observational Health Data Sciences and Informatics. Available from: https://github.com/OHDSI [Accessed 5th March 2019].
23. SAS. SAS analytics software solutions. Available from: https://www.sas.com/en_us/solutions/analytics.html [Accessed 5th March 2019].
24. MacLennan S, Williamson PR, Bekema H, Campbell M, Ramsay C, N'Dow J, et al. A core outcome set for localised prostate cancer effectiveness trials. BJU Int. 2017;120(5B):E64-E79.
25. Martin NE, Massey L, Stowell C, Bangma C, Briganti A, Bill-Axelson A, et al. Defining a standard set of patient-centered outcomes for men with localized prostate cancer. Eur Urol. 2015;67(3):460-7.
26. Morgans AK, van Bommel AC, Stowell C, Abrahm JL, Basch E, Bekelman JE, et al. Development of a Standardized Set of Patient-centered Outcomes for Advanced Prostate Cancer: An International Effort for a Unified Approach. Eur Urol. 2015;68(5):891-8.
27. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015;350:g7647.
28. Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. J Clin Epidemiol. 2018;96:84-92.
29. Williamson PR, Altman DG, Bagley H, Barnes KL, Blazeby JM, Brookes ST, et al. The COMET Handbook: version 1.0. Trials. 2017;18(Suppl 3):280.
30. Bruinsma SM, Roobol MJ, Carroll PR, Klotz L, Pickles T, Moore CM, et al. Expert consensus document: Semantics in active surveillance for men with localized prostate cancer - results of a modified Delphi consensus procedure. Nat Rev Urol. 2017;14(5):312-22.
31. HARMONY. Big Data To Enable Better And Faster Treatments For Patients With Hematological Malignancies. Available from: https://www.harmony-alliance.eu/ [Accessed 5th March 2019].
32. Scheufele E, Aronzon D, Coopersmith R, McDuffie MT, Kapoor M, Uhrich CA, et al. tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform. AMIA Jt Summits Transl Sci Proc. 2014;2014:96-101.
33. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11(3):333-7.
34. Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, et al. Extracting insights from the shape of complex data using topology. Sci Rep. 2013;3:1236.
35. Le Cao KA, Rohart F., Gonzalez .I, Dejean S., Gautier B., Bartolo F., et al. mixOmics: Omics Data Integration Project. 2016. p. R package version 6.1..
36. Lowe WL, Jr., Reddy TE. Genomic approaches for understanding the genetics of complex disease. Genome research. 2015;25(10):1432-41.
37. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A. caret: Classification and Regression Training. 5.15-044 ed2012.
38. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med. 2001;20(6):825-40.
39. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002;21(11):1559-73.
40. Reimand J, Arak T, Vilo J. g:Profiler--a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. 2011;39(Web Server issue):W307-15.
41. Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic acids research. 2017;45(D1):D877-D87.
42. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic acids research. 2017;45(D1):D362-D8.
43. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. York S-VN, editor2009.

44. Alberts AR, Schoots IG, Bokhorst LP, Drost FH, van Leenders GJ, Krestin GP, et al. Characteristics of Prostate Cancer Found at Fifth Screening in the European Randomized Study of Screening for Prostate Cancer Rotterdam: Can We Selectively Detect High-grade Prostate Cancer with Upfront Multivariable Risk Stratification and Magnetic Resonance Imaging? European urology. 2017.

45. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W, 1 GWP. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. Value Health. 2017;20(7):858-65.

46. EUR-Lex. Access to European Union law. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 [Accessed 5th March 2019].

Box 1: PIONEER Research Objectives

PIONEER aims to optimise diagnosis and therapeutic management of prostate cancer patients across different stages of the disease and across multiple geographies by delivering valuable insights from clinical and real-world data and sharing best practices (all WPs).

- To improve disease understanding and deliver a core set of clinically relevant standardised prostate cancer-related outcomes (WP2 with WP3, WP4 and WP5).
- To develop a large and harmonised repository of prostate cancer data that can be used to improve evidence-based decision-making for all prostate cancer patients and enable a wide variety of data re-use scenarios (WP4 with WP3 and WP5).
- To provide unique tools for standardisation and analysis of complex prostate cancer data sets from a variety of sources, using different data models and different terminology, whilst taking into account different layers of information (e.g. genomic, transcriptomics, etc.) (WP3 and WP5, with WP4 contributing).
- To raise awareness, dissemination and widespread implementation of PIONEER results (WP6 and WP7 with all WPs).
- To address the barriers related to data sharing and data protection (WP8 with all WPs).