

# 1 ESSAY

## 2 Introducing PIONEER: a project to harness big data in prostate cancer 3 research

4  
5 Muhammad Imran Omar<sup>a</sup>, Monique J. Roobol<sup>b</sup>, Maria J. Ribal<sup>c</sup>, Thomas Abbott<sup>d</sup>, Paul-Michael  
6 Agapow<sup>e</sup>, Sonia Araujo<sup>f</sup>, Alex Asiiimwe<sup>g</sup>, Charles Auffray<sup>h</sup>, Irina Balaur<sup>h</sup>, Katharina Bayer<sup>i</sup>, Chiara  
7 Bernini<sup>j</sup>, Anders Bjartell<sup>k</sup>, Alberto Briganti<sup>l</sup>, John-Edward Butler-Ransohoff<sup>g</sup>, Riccardo Campi<sup>c</sup>, Marinel  
8 Cavelaars<sup>m</sup>, Bertrand De Meulder<sup>h</sup>, Zsuzsanna Devecseri<sup>n</sup>, Marc Dietrich Voss<sup>n</sup>, Konstantinos  
9 Dimitropoulos<sup>c</sup>, Susan Evans-Axelsson<sup>k</sup>, Billy Franks<sup>d</sup>, Louise Fullwood<sup>o</sup>, Denis Horgan<sup>l</sup>, Emma Jane  
10 Smith<sup>c</sup>, Amit Kiran<sup>d</sup>, Kati Kivinummi<sup>p</sup>, Mark Lambrecht<sup>q</sup>, Doron Lancet<sup>r</sup>, Peter Lindgren<sup>s</sup>, Sara  
11 MacLennan<sup>a</sup>, Steven MacLennan<sup>a</sup>, Maria Manuela Nogueira<sup>h</sup>, Fredrik Moen<sup>s</sup>, Maxim Moinat<sup>m</sup>, Kishore  
12 Papineni<sup>d</sup>, Christian Reich<sup>f</sup>, Kristin Reiche<sup>t</sup>, Stijn Rogiers<sup>q</sup>, Claudio Sartini<sup>g</sup>, Kees van Bochove<sup>m</sup>, Femke  
13 van Diggelen<sup>u</sup>, Mieke Van Hemelrijck<sup>i</sup>, Hein Van Poppel<sup>c</sup>, Jihong Zong<sup>g</sup>, James N'Dow<sup>c</sup> and The PIONEER  
14 Consortium\*

### 15 Affiliations:

16 a: Academic Urology Unit, University of Aberdeen, Aberdeen, United Kingdom

17 b: Erasmus MC, Rotterdam, The Netherlands

18 c: Guidelines Office, European Association of Urology, Arnhem, The Netherlands

19 d: Astellas, The Netherlands

20 e: Imperial College London, London, United Kingdom

21 f: IQVIA

22 g: Bayer AG, Berlin, Germany

23 h: Association EISBM, France

24 i: Translational Oncology and Urology Research, King's College London, London, United Kingdom

25 j: European Alliance for Personalised Medicine (EAPM), Belgium

26 k: Department of Translational Medicine, Lund University, Lund, Sweden

27 l: Department of Urology and Division of Experimental Oncology, Urological Research Institute, Vita-  
28 Salute San Raffaele University, IRCCS San Raffaele Scientific Institute, Milan, Italy

29 m: The Hyve, Utrecht, The Netherlands

30 n: Sanofi, France

31 o: Pinsent Masons, United Kingdom

32 p: Tampere University (TAU), Tampere, Finland

33 q: SAS, Belgium

34 r: Weizmann Institute, Rehovot, Israel

35 s: The Swedish Institute for Health Economics (IHE), Stockholm, Sweden

36 t: Fraunhofer IZI, Leipzig, Germany

37 u: ttopstart, Utrecht, The Netherlands

38  
39 \*The PIONEER Consortium include listed authors and the following collaborators: Emelie Andersson, Heidi Arala, Anssi  
40 Auvinen, Chris Bangma, Danny Burke, Antonella Cardone, Joaquin Casariego, Guido Cuperus, Saeed Dabestani, Francesco  
41 Esperto, Nicola Fossati, Adam Fridhammar, Giorgio Gandaglia, Delila Gasi Tandefelt, Friedemann Horn, Johannes Huber,  
42 Jonas Hugosson, Henkjan Huisman, Michelle Jones, Andreas Josefsson, Olavi Kilkku, Markus Kreuz, Michael Lardas, Joe  
43 Lawson, Florence Lefresne, Stephane Lejeune, Elaine Longden-Chapman, Gordon McVie, Lisa Moris, Teemu Murtola, Charlie  
44 Nicholls, Karl H. Pang, Katie Pascoe, Marta Picozzi, Karin Plass, Pasi Pohjanjousi, Matthew Reaney, [Sebastiaan Remmers](#), Paul  
45 Robinson, Jack Schalken, Max Schravendeel, Thomas Seisen, Angela Servan, Kirill Shiranov, Robert Snijder, Nesrine Taibi,  
46 Kirsi Talala, Derya Tilki, Thomas Van den Broeck, Zdravko Vassilev, Olli Voima, Eleni Vradi, Reg Waldeck, Ward Weistra, Peter-  
47 Paul Willemse, Manfred Wirth, Russ Wolfinger, Nazanin Zounemat Kermani

48 **Abstract:**

49 PIONEER (Prostate Cancer Diagnosis and Treatment Enhancement through the power of big  
50 data in Europe) is a European network of excellence for big data in prostate cancer, consisting  
51 of 32 private and public stakeholders from 9 countries across Europe. Launched by the  
52 Innovative Medicines Initiative 2 and part of the Big Data for Better Outcomes Programme  
53 (BD4BO), the overarching goal of PIONEER is to provide high-quality evidence on prostate  
54 cancer management by unlocking the potential of big data.

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

67 **Introduction**

68

69 Prostate cancer is the second most common cancer in men by incidence in Europe, with  
70 450,000 new cases diagnosed in 2018. Prostate cancer incidence varies five-fold, with the  
71 highest incidence in Northern and Western Europe, and the lowest in Central and Eastern  
72 Europe. The estimated incidence is highest in Ireland (189.3 per 100,000), whereas Albania  
73 (37 per 100,000) and Romania (47.2 per 100,000) have the lowest incidence (1). In 2018, the  
74 estimated numbers of death of prostate cancer were 107,300 for Europe (40 European  
75 countries), and 81,500 for 28 member countries of the European Union (1). Total annual  
76 estimated costs for treatment of prostate cancer in the first year following diagnosis is  
77 approximately €117 million in the UK. The figure is two- to three-fold higher in France and  
78 Germany (2). This economic burden associated with prostate cancer is predicted to  
79 dramatically increase in the coming years due to aging of the population, as around 85% of  
80 all cases of prostate cancer are diagnosed in men over the age of 65 years (1, 3, 4). Despite  
81 these numbers, up to now the level of funding for research is relatively low. For example, in  
82 2018/2019, Cancer Research UK spend £13 million on prostate cancer research out of their  
83 total annual budget of £442 million (5). Therefore, progress made in prostate cancer research  
84 is limited when compared to other major cancer types. (1) For example, mortality statistics of  
85 Cancer Research UK indicate the mortality rate of breast cancer has been steadily declining,  
86 while the prostate cancer mortality rate is still on the rise (5).

87

88 Currently, several critical questions remain unresolved regarding the screening, diagnosis and  
89 treatment of prostate cancer patients, relating to various observations in prostate cancer  
90 epidemiology. First, prostate cancer incidence is variable across different European countries  
91 (37 to 189 per 100,000) (1). The differences in incidence rates of different racial and ethnic  
92 background confirms the involvement of genetic factors. However, environmental factors  
93 may also be implicated as the differences are also observed among men of the same genetic

94 heritage who live in different European countries. Furthermore, inequalities in prostate  
95 cancer survival are also observed across the European Union. Estonia and Latvia have the  
96 highest mortality rates (37.3 per 100,000 and 35.7 per 100,000 respectively), whereas the  
97 mortality rates are the lowest in Spain and Italy (13.2 per 100,000 and 10.7 per 100,000  
98 respectively) (1).

99

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

115

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

121

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

134

### 135 **PIONEER's vision**

136

137 The vision of PIONEER is to transform the management and clinical practice of prostate cancer

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

150

### 151 **Objectives of PIONEER**

152

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

156

### 157 **Approach and methodology**

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

169 PIONEER has developed a dual approach, in order to use prostate cancer big data to develop  
170 an outcome-driven, value-based, and patient-centric healthcare system. First, PIONEER will  
171 identify critical evidence gaps in prostate cancer by combining the knowledge of academic  
172 and industry professionals and patients, thus enabling to focus the PIONEER working plan on  
173 a consensus list of research priorities and questions. Then, PIONEER will integrate, analyse,  
174 standardise and harmonise existing data from high quality and multidisciplinary data sources  
175 from prostate cancer patients across different stages of the disease into a single data  
176 platform (15, 16). To achieve this, PIONEER will use readily available, successful workbench  
177 and tools, such as tranSMART, OHDSI and the SAS open Platform, based on suitable data  
178 harmonisation techniques (OMOP Common Data Model) and advanced analytical methods.  
179 The advanced analytical methods may include machine learning, predictive modelling, multi-  
180 omics data integration methods, data visualisations as developed by The Hyve in the IMI1

181 funded project EMIF (the European Medical Information Framework) (17) and by the  
182 European Institute for Systems Biology and Medicine (EISBM (18)) and the Data Science  
183 Institute at Imperial College London (DSI-ICL (19)) in the IMI1 U-BIOPRED (20) and eTRIKS (21)  
184 projects.

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

## 190 **Prioritisation of the most important questions in the field of prostate cancer**

191  
192 The EAU Prostate Cancer Guideline panel and other prostate cancer Key Opinion Leaders  
193 were contacted to identify the most important questions in the field of prostate cancer. Forty-  
194 four viable questions were identified. Afterwards, the PIONEER consortium performed a  
195 prioritisation survey among two stakeholder groups: healthcare professionals including  
196 pharmaceutical companies and prostate cancer patients.

197  
198 In total, 73 healthcare professionals and 57 patients participated in round one of the surveys.  
199 The results were analysed by calculating the percentage of respondents scoring each question  
200 as not important, important or critically important.. Twelve additional questions were  
201 proposed during the first round. For the second round the patients' surveys were also  
202 translated into French, German, Italian and Spanish. 49 healthcare professionals and 169  
203 patients (including 53 English; 19 French; 31 German; 53 Italian; 13 Spanish) participated in  
204 round two of the surveys. These 56 questions (44 questions from round one and 12 additional  
205 questions from round two) were then re-ordered according to the highest percentage for  
206 "critically important". The questions covered all stages of prostate cancer focusing on various  
207 aspects of the condition including screening, diagnosis, risk stratification including the  
208 genomic profile, treatment, and complications of treatment. The detailed results will be  
209 presented in a separate publication, but in meantime are being used to inform PIONEER  
210 consortium, so that the stakeholder groups' priorities are met in an accountable and  
211 transparent way.

212

213

## 214 **WP1: Project management and administration**

215

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

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

229

## 230 **WP2: Disease understanding and outcome definition**

231

232 The aim of WP2 is to develop standardised definitions and measurements of prostate cancer  
233 outcomes and diagnostic, predictive, prognostic, and therapeutic factors (DPPTs) across the different  
234 stages of prostate cancer care, and to consider the opinions of key stakeholders in this process (**Box**  
235 **1: PIONEER Research Objectives**).

236

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

253

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

258

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

269

270 In addition to assessing published literature in our systematic reviews, we will also evaluate  
271 the data collected in the different data resources of PIONEER. This process will result in a

272 structured database of verbatim outcome names, definitions and measures. The outcomes  
273 database will be categorised according to the generic Williamson Clarke taxonomy (28), with  
274 additional prostate cancer specific outcomes and definitions provided (**elaborated below in**  
275 **WP4: Data platform**). This will structure and homogenise the available COSs.

276

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

282

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

293

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

308

### 309 **WP3: Data access and sources**

310

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

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

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

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

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

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

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

#### 354 **WP4: Data platform**

355

356 PIONEER WP4 will develop a pan-European data-sharing platform and adopt a two-pronged  
357 approach to address the project needs: a) a platform that can access population-based



358 registry data such as electronic health records, and b) a platform that can handle rich clinical  
359 and omics data for translational analysis by WP5 (**elaborated below in WP5: Data analytics**).  
360 To achieve this, the project will build upon and use approaches developed in a number of  
361 other IMI projects, such as IMI1 EMIF (17) and IMI1 eTRIKS (21).

362

363 For data integration and analysis of longitudinal prostate cancer registries, PIONEER will use  
364 the OMOP and OHDSI (22) technology, while for cohort studies that also include omics data  
365 besides deep phenotypic and clinical data, the tranSMART (32) technology will be used (**Figure**  
366 **2**). Components from both technologies are still under development.

367

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

373

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

378

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

384

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

### 391 **WP5: Data analytics**

392

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

395 The team in WP5 will provide a unique toolkit of standard and cutting-edge analytical  
396 methods for the analysis of big data, both from open-source and industry-developed methods  
397 (**Box 1: PIONEER Research Objectives**). Research questions and core outcome sets have been

398 identified in PIONEER's survey conducted by WP2. Each of the research questions that  
399 PIONEER will tackle will require different tools and analytic workflows that will be provided  
400 by WP5, through the centralised tranSMART omics platform built by WP4 (**Figure 3**).

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

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

432 The list of analytical tools that are expected to be of use in PIONEER is still being refined.  
433 However, there will be a heavy need for predictive modelling and machine learning (random  
434 forests, linear modelling, support-vector machines, partial least square regressions).  
435 Visualisation will be provided both at the level of the data platform (either in tranSMART or  
436 OHDSI) directly and with dedicated software included in the SAS suite and in R packages. We  
437 will also monitor and make use of developments in the broader computational systems  
438 biology community as they become available to use. High-performance computing, bringing  
439 big data analytics capabilities when needed to answer PIONEER research questions will be

440 provided through a SPARK infrastructure hosted at the DSI-ICL. Finally, WP5 will make use of  
441 the developments, insights and experience of other research projects through partnerships,  
442 research seminars and the projects members' experience, either from IMI projects [eTRIKS,  
443 B4B (Brains for Brain), EMIF, parallel BD4BO IMI projects] or from the industry partners'  
444 internal knowledge and developments.

#### 445 **WP6: HTA regulator**

446  
447 Through WP6, the PIONEER project will seek to develop, and also validate, a framework for  
448 innovative technologies in prostate cancer using real-world evidence (**Box 1: PIONEER**  
449 **Research Objectives**). The latter involves using various health data in real time to help  
450 healthcare professionals make better and quicker decisions. Real-World data has been  
451 defined as "an umbrella term for different types of healthcare data that are not collected in  
452 conventional randomized controlled trials... including patient data, data from clinicians,  
453 hospital data, data from payers and social data" (45).

454  
455 Many HTA and payer groups think of real-world evidence as having much potential, but  
456 alignment is still necessary. PIONEER will work with such bodies as well as regulators to  
457 establish minimum evidence requirements while identifying, at an early stage, potential  
458 uncertainties requiring extra data. On top of this, PIONEER will seek to develop reference  
459 models for use in economic evaluations and, as a key objective, will explore whether it can  
460 develop a core set of reference models for different stages of prostate cancer, or an  
461 overarching modelling framework. This is necessary in order to explore the impact of new  
462 technologies at single points along the pathway, as well as looking at treatment sequences,  
463 as the disease progresses through its multiple stages.

464  
465 Effective evaluation of the medical, social, economic and ethical issues of products in a  
466 systematic, transparent, unbiased, robust manner will promote safe, effective, health policies  
467 that are patient-focused and obtain best value - whether at the time of launch or their usage  
468 in real-life circumstances. Adapted tools and openness to evidence produced by methods  
469 other than classic RCTs will be helpful. In the case of adaptive clinical trials, real-world  
470 evidence is crucial. Medical Adaptive Pathways to Patients, known as MAPPs, have been  
471 tested in a European Medicines Agency (EMA) project, and will be used in PIONEER.

472  
473 MAPPs are described as a prospectively planned process, starting with the early authorisation  
474 of a medicine in a restricted patient population, followed by iterative phases of evidence  
475 gathering and adaptations of the marketing authorisation to expand access to the medicine  
476 to broader patient populations. The keywords here are the 'iterative phases of evidence  
477 gathering', which should use real-world evidence to detect patient responses to new  
478 therapies in a real-time setting.

479 Meanwhile, many of those mobile applications that we are all now using are essentially  
480 gathering real-world evidence on a daily basis. The more advanced health applications can  
481 provide this while also running simultaneous comparative efficacy trials against existing  
482 therapies. These applications could also solve issues surrounding data interoperability, given  
483 that a data standard is already in place when using iOS or Android platforms. In theory, these

484 real-time datasets could then be sent, for example, as standard XML files to any internet  
485 database in Europe and beyond.

486

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

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

#### 494 **WP7: Dissemination and communication**

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

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

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

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

520 Identification of PIONEER target audience is a key step for successful dissemination of the  
521 project outcomes. Successful identification and engagement of all relevant stakeholders  
522 could be a potential challenge. Knowing our target audiences involves knowing the specific  
523 needs of the individual audience and not just the message we want to convey. To overcome

524 communication barriers, it is important to determine the medium through which PIONEER  
525 will communicate with different target audiences and the timing of message delivery.  
526 Information needs to be of good quality, timely, contextually relevant and appropriate to the  
527 intended audience. Furthermore, the early involvement of all relevant stakeholders is a key  
528 enabler: being actively involved in the design and prioritisation of the research questions  
529 addressed throughout the project will help greatly in ensuring ongoing stakeholder  
530 engagement enabling PIONEER to meet its objectives.

#### 531 **WP8: Legal, ethical issues and governance**

532

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

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

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

564

#### 565 **Planned outcomes of PIONEER**

566

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

573

574 • Consensus on the most important prostate cancer outcomes (WP2: Disease  
575 understanding and outcome definition)

576 • Identification of critical evidence gaps in prostate cancer (as detailed above under:  
577 Prioritisation of the most important questions in the field of prostate cancer

578 • Standardisation of outcome definition and outcome measures outcomes

579 • New insights on improved stratification

580 • Improved standardized care pathways with known better predictable outcomes

581

### 582 **Challenges in PIONEER**

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

595

### 596 **Future directions**

597

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

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

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

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

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

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

616

CONFIDENTIAL

## 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 5<sup>th</sup> 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 24<sup>th</sup> 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 5<sup>th</sup> 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](http://www.imi.europa.eu) [Accessed 5<sup>th</sup> March 2019].
15. BD4BO. Big Data for Better Outcomes. Available from: <http://bd4bo.eu> [Accessed 5<sup>th</sup> 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 5<sup>th</sup> March 2019].
18. EISBM. European Institute for Systems Biology and Medicine. Available from: <http://www.eisbm.org/> [Accessed 5<sup>th</sup> March 2019].
19. London IC. Data Science Institute. Available from: <https://www.imperial.ac.uk/data-science/about-the-institute/> [Accessed 5<sup>th</sup> 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 5<sup>th</sup> March 2019].
21. eTRIKS. eTRIKS Harmonisation Services. Available from: <https://www.etriks.org/etriks-harmonisation-services/> [Accessed 5<sup>th</sup> March 2019].



22. OHDSI. Observational Health Data Sciences and Informatics. Available from: <https://github.com/OHDSI> [Accessed 5<sup>th</sup> March 2019].
23. SAS. SAS analytics software solutions. Available from: [https://www.sas.com/en\\_us/solutions/analytics.html](https://www.sas.com/en_us/solutions/analytics.html) [Accessed 5<sup>th</sup> 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, Gherzi 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 5<sup>th</sup> March 2019].
32. Scheufele E, Aronzon D, Coopersmith R, McDuffie MT, Kapoor M, Uhrich CA, et al. transSMART: 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 5<sup>th</sup> March 2019].

CONFIDENTIAL

### 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).