Protocol

INTERNATIONAL JOURNAL OF SURGERY PROTOCOLS

OPEN

# Research protocol for an observational health data analysis to assess the applicability of randomized controlled trials focusing on newly diagnosed metastatic prostate cancer using real-world data: PIONEER IMI's "big data for better outcomes" program

Giorgio Gandaglia, MD[a,*], Francesco Pellegrino, MD[a], Bertrand De Meulder, MSc[c], Ayman Hijazy, PhD[c], Thomas Abbott, PhD, MBA[d], Asieh Golozar, PhD[e], Rossella Nicoletti, MD[b], Juan Gomez-Rivas, PhD[f], Carl Steinbeisser, MSc[g], Susan Evans-Axelsson, MSc[h], Alberto Briganti, PhD[a], James N'Dow, MD[i]

**Background:** Metastatic prostate cancer (PCa) constitutes ~5% of all new PCa diagnoses in Western countries. For most cases, primary consideration should be given to systemic therapies as the first-line approach based on evidence from randomized controlled trials (RCTs). Despite the importance of RCTs as the pinnacle of evidence in modern medicine, concerns have been raised about their applicability to real-life scenarios. These trials often feature participants who are younger with better performance statuses and prognoses compared to their real-world counterparts. The PIONEER project falls under the Innovative Medicine Initiative's (IMI) "Big Data for Better Outcomes" initiative, aimed at revolutionizing PCa care in Europe. The central focus lies in improving cancer-related outcomes, enhancing health system efficiency, and elevating the quality of health and social care. This study endeavours to evaluate the generalizability of RCT findings concerning newly diagnosed metastatic PCa.

**Methods:** A systematic review of the literature will be conducted to compile patient characteristics from RCTs addressing this subject within the past decade. To create a real-world benchmark, patients with recently diagnosed metastatic PCa from a network of population-based databases will serve as a comparison group. The objective is to assess the applicability of RCT results in two ways. First, a comparison will be made between the characteristics of patients with newly diagnosed metastatic PCa enroled in RCTs and those with the same condition included in our databases which might represent the real-world setting. Second, an evaluation will be undertaken to determine the proportion of real-world patients with newly diagnosed metastatic PCa who meet the criteria for RCT enrolment. This study will rely on extensive observational data, primarily sourced from population-based registries, electronic health records, and insurance claims data. The study cohort is established upon routinely gathered healthcare data, meticulously mapped to the Observational Medical Outcomes Partnership Common Data Model.

**Keywords:** Metastatic prostate cancer, prostatic cancer, randomized controlled trials

## Rationale and background

Prostate cancer (PCa) is the second most common solid tumour and the second leading cause of cancer deaths in men worldwide[1,2]. The widespread adoption of early detection strategies based on prostate-specific antigen (PSA) led to an increase in PCa incidence and a stage migration phenomenon, with more indolent PCa being detected and fewer cases of metastatic PCa being diagnosed in the USA and Europe[3]. However, metastatic PCa still accounts for more than 5% of all new PCa diagnoses in Western countries[4]. The updated guidelines recommend offering androgen deprivation therapy (ADT) combined with androgen receptor pathway inhibitors and/or

[a]Division of Oncology/Unit of Urology, Soldera Prostate Cancer Lab, URI, IRCCS San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Milan, [b]Department of Experimental and Clinical Biomedical Science, University of Florence, Florence, Italy, [c]Association EISBM, Vourles, France, [d]Astellas Pharma, Inc., Northbrook, IL, [e]OHDSI Center, Northeastern University, Boston, MA, [f]Department of Urology, Hospital Clínico San Carlos, Madrid, Spain, [g]Collaborate Project Management, Munich, [h]Bayer AG, Berlin, Germany and [i]Academic Urology Unit, University of Aberdeen, Scotland, UK

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

*Corresponding author. Address: Division of Oncology/Unit of Urology, Soldera Prostate Cancer Lab, URI, IRCCS San Raffaele Hospital, Urological Research Institute, Milan, Italy; Tel.: +39 02 2643 4066; fax: +39 02 2643 7298. E-mail: gandaglia.giorgio@hsr.it (G. Gandaglia).



International Journal of Surgery Protocols (2024) 28:64–72

Received 7 October 2023; Accepted 16 March 2024

Published online 15 April 2024

http://dx.doi.org/10.1097/SP9.0000000000000024

chemotherapy as first-line therapy for these patients[5,6]. These recommendations are based on the results of randomized controlled trials (RCTs) which demonstrated a survival benefit of this therapeutic approach[7–18]. Of note, RCTs represent the highest hierarchical level of evidence based on a single experiment[19]. These studies are designed to determine the efficacy of an intervention under idealized and controlled circumstances including strict adherence to structured protocols, the use of restrictive inclusion and exclusion criteria, and patient randomization[20]. For RCTs to be clinically useful, their results must also be applicable to a defined group of patients in a specific therapeutic setting in the real-life scenario; this is known as external validity (or generalizability)[20]. The generalizability of RCTs has been called into question[20–22]. Indeed, a literature review found that less than 40% of cancer patients would be eligible for RCTs[21]. Indeed, patients included in RCTs are often younger, healthier and with a less aggressive cancer than those with the same disease in the real-world scenario[20–22]. Moreover, RCTs are typically carried out in high-volume centres, which might not reflect what is happening in the daily clinical practice[20–22]. This hypothesis has been tested and demonstrated only for few cancers[21], and never for PCa. In the face of such a paucity of data, the PIONEER Consortium, which is a novel project of the Innovative Medicine Initiative's (IMI's) "Big Data for Better Outcomes" program with the mission to transform PCa care with particular focus on improving cancer-related outcomes, health system efficiency and the quality of health and social care across Europe. The EAU Prostate Cancer Guideline panel and other PCa Key Opinion Leaders were consulted to propose the most critical questions in the field of PCa to be answered using big data. Through this process, 44 key questions were identified. Afterwards, the PIONEER consortium conducted a two round Delphi survey to build consensus between the two stakeholder groups: healthcare professionals (including representatives from pharmaceutical companies) and PCa patients. Respondents were asked to consider what impact answering the proposed questions would have on better diagnosis and treatment outcomes for PCa, while scoring these questions [on a scale of 1 (not important) to 9 (critically important)]. The results were analyzed by calculating the percentage of respondents scoring each question as not important (score 1–3), important (score 4–6) or critically important (score 7–9). A modified Delphi Methods was then adopted for this prioritization process to build consensus among the participants. In the second round, participants were shown a summary of the percentage of other participants' (patients and healthcare professionals) who considered the question "critically important" in round one. Among the identified clinical questions, the following ones aimed at evaluating the generalizability of RCTs on newly diagnosed metastatic PCa were prioritized: "Can we integrate data coming from randomized trials into population-based and prospective cancer registries?" and "Are results obtained using currently available data sources generalizable to all PCa patients?"[23].

## Objectives

This study aims to evaluate the external validity of RCTs on newly metastatic PCa. In detail, this study aims:

(1) To compare the characteristics of patients with newly diagnosed metastatic PCa who were enroled in RCTs with those of patients with this disease in the real-world setting;
(2) To assess the proportion of real-world patients with newly diagnosed metastatic PCa who would have been eligible for RCT inclusion following the application of RCT selection criteria.

### HIGHLIGHTS

- Most patients with metastatic prostate cancer (PCa) at diagnosis are typically managed with systemic therapy based on findings from randomized controlled trials (RCTs).
- The generalizability of RCTs has been called into question.
- Our study will evaluate the generalizability of RCTs findings concerning newly diagnosed metastatic PCa.
- We aim at identifying areas for further research and improve the interpretation and application in daily clinical practice of the RCTs findings.

## Hypotheses

The null hypothesis of this study is that there is no difference in the characteristics of patients with newly metastatic PCa enroled in RCTs and those with the same condition in real-world settings.

## Methods

### *Data sources*

The study will rely on large observational data, namely population-based registries, electronic health records (EHR) and insurance claims data (Table 1). The data will be analyzed using a federated model, where the data remain with the data custodians and only the analysis results are shared and published. Case series will not be considered.

### *Study design*

The study will take place in two steps. First, we will perform a systematic review of the literature to identify studies reporting the results of RTCs on newly diagnosed metastatic PCa. We will collect demographics and clinical characteristics of the patients included in the RCTs and selection criteria used by each RCTs reviewing the original protocol published with the full manuscript or in ClinicalTrials.gov. Second, we will identify patients with newly metastatic PCa in an observational cohort based on routinely collected healthcare data which has been mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Patients' demographics and clinical characteristics at or prior to index date (defined below) and treatments of these individuals at or after their index date will be described (clinical characterization). Then, we will compare demographics and clinical characteristics of these patients with the characteristics of patients who were enroled in the selected RCTs. Finally, we will assess the proportion of patients in out cohort with newly diagnosed metastatic PCa who would have been eligible for RCT inclusion following the application of RCTs' selection criteria.

**Table 1**

**Data sources formatted to the OMOP CDM with prostate cancer patients to be included in the PIONEER Study-a-thon as of April 2021[a].**

| Source full name | Country code | Data provenance | Source short name | Patient count | History | Patient type | Data collection |
|---|---|---|---|---|---|---|---|
| Optum de-identified Electronic Health Record Dataset | USA | EMR | Optum EHR - EMR, US | 96 million | -2006 | EHR/Privately Insured | Optum© de-identified Electronic Health Record Dataset represents Humedica's Electronic Health Record data a medical records database. The medical record data includes clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical Notes using natural language processing (NLP). |
| CDW Bordeaux University | FRA | EHR | Bordeaux | ~6700 | | | Electronic Health Records<br>- 1.8M persons<br>- Malignant prostate cancer<br>- 6700 persons (ICD-10 code: C61) •40% dead (in or out-hospital death)<br>- Hospital treatments: Surgery, Radiation therapy, Hormone therapy, Chemotherapy, Active surveillance /Watchful waiting |
| Netherlands Cancer registry | NLD | | IKNL | 295 000 | | | The NCR compiles clinical data of all individuals newly diagnosed with cancer in the Netherlands. Cancer registration clerks register newly diagnosed cancer patients since 1989 on a national basis. Over the past 30 years, this registry has provided clinicians and researchers with a wealth of clinical data (e.g. stage and primary treatment) on cancer patients of all ages. See https://iknl.nl/en for more information. |
| MAITT (University of Tartu and STACC) | EST | EHR, insurance claims, prescriptions | MAITT | ~17 000–18 000 | 2012–2019 | | 10% random sample from Estonian population<br>- All 2012–2019 records of the sample from three data sources:<br>o Digital Prescription database of Estonia - > 99% of all prescriptions<br>o Insurance claims from national Health Insurance Fund (EHIF) -95% of Estonian people are insured by EHIF<br>o EHR from national centre Health Information System (HWISC) - central repository for electronic health records (EHR)<br>- Counts: ~150 000 persons all together and ~1700 patients with prostate cancer (ICD-10 diagnosis code C61) |
| Medaman | NDL | | Medaman | 3130 | Pathology, Procedures, selected lab results from 2016–present. Medications from 2020–present | | "Co-ordinating and supporting Belgian hospitals during standard real-world data coding and bi-annual official reporting to Federal Ministry of Health<br>Organising hospital-based Clinical Documentation Improvement initiatives<br>Data analysis<br>Data visualization (Tableau)<br>Hospital data:<br>Emergency department |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | One-day care<br>Hospital stay" |
| Durham Veterans Affair Medical Center | USA | | DVAMC | millions | | | VA OMOP data reflects the national Department of Veterans Affairs healthcare system, which is the largest integrated provider of medical and mental health services in the United States. Care is provided at 170 VA Medical Centers and 1063 outpatient sites serving more than 9 million enroled Veterans each year. |
| Epic Legacy Columbia University Irvine Medical Center (CUIMC) MERGE | USA | | CUIMC | 6 666 613 | | | General practice electronic health records, outpatient specialist electronic health records, inpatient hospital electronic health records, hospital billing/summary |
| TUFTS Medical Center | USA | EHR, Tumour Registry, State vital statistics. | TRDW | ~1 000 000 | 2007–present | Adult and paediatric patients who receive care in inpatient or outpatient settings that are part of Tufts MC. | Data are collected from multiple EHR sources including a oncology-specific EHR, and fused with information on the same patients from the CoC-accredited Tufts Tumour Registry (using mappings from the NAACCR format to OMOP) and with current death data from the Massachusetts Registry of Vital Records and Statistics. NLP is used to collect and standardize concepts from unstructured sources in the EHR on a variety of symptoms, measurements (e.g. echocardiogram-derived measures of heart function). |
| Clinical Practice Research Datalink (CPRD) | USA | | CPRD | 58 603 | | | Electronic health records from primary care. Data includes conditions, observations, measurements, and procedures that the general practitioner is made aware of in additional to any prescriptions as prescribed by the general practitioner. |
| MarketScan | USA | | MarketScan | | | | Commercial Claims and Encounters, Medicare Supplements captures person-specific clinical utilization, expenditures, and enrolment across inpatient, outpatient, prescription drug, and carve-out services from a selection of large employers, health plans, and government and public organizations. |
| OPTUM claims | USA | EMR | OPTUM claims | 651 765 | | | Optum© de-identified Electronic Health Record Dataset represents Humedica's Electronic Health Record data a medical records database. The medical record data includes clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical notes using natural language processing (NLP). |
| IQVIA OpenClaims | USA | | OpenClaims | | | | A United States database of open, pre-adjudicated claims from January 2013 to May 2020. Data are reported at anonymized patient level collected from office-based physicians and specialists via office management software and clearinghouse switch sources for the purpose of reimbursement. A subset of medical claims data have adjudicated claims. |

**Table 1**

**(*Continued*)**

| Source full name | Country code | Data provenance | Source short name | Patient count | History | Patient type | Data collection |
|---|---|---|---|---|---|---|---|
| IQVIA OncoEMR | USA | | OncoEMR | | | | Oncology EMR US is comprised of anonymized patient records collected from electronic health records in oncology specialty hospital setting. Data coverage includes 1.36M patients, 14.8K providers and 552 care sites with outpatient visit. Dates of service include from 1997 through present. Observation time is defined by the first and last consultation dates. Drugs are captured as prescription records with product, quantity, dosing directions, strength, indication and date of consultation. |
| IQVIA PharMetrics Plus | USA | | Pharmetrcis Plus | | | | PharMetrics Plus (PMTX+) US is comprised of anonymized patient records collected from adjudicated claims (accepted and paid by the payer). Data coverage includes 145M patients across inpatient, outpatient, ER, Pharmacy, Laboratory, Ambulance, Home, Telehealth and Non-hospital visits, covering 12.8% of the national population. Dates of service include from 2006 through present. Observation time is defined by the first and last consultation dates. Drugs are captured as prescription records with product, quantity, dosing directions, strength, indication and date of consultation. |
| Information System for Research in Primary Care (SIDIAP) | ESP | | SIDIAP | 26 000 | | | The Information System for Research in Primary Care (SIDIAP; www.sidiap.org) is a primary care records database that covers ~80% of the population of Catalonia, North-East Spain. Healthcare is universal and tax-payer funded in the region, and primary care physicians are gatekeepers for all care and responsible for repeat prescriptions. |

[a]This does not represent the final list and more databases might be included if they are deemed suitable for the purpose of the study.

EMR, electronic medical records; IKNL, Integraal Kankercentrum Nederland; NCR, Netherlands Cancer Registry; DVAMC, durham veterans affair medical center; TRDW, tufts research data warehouse

### RCT identification and evaluation

A systematic review of the literature will be conducted using the PubMed/MEDLINE, Cochrane library's Central, EMBASE, and Scopus databases. We will search from January 2013 up to 1 June 2023. This time frame was chosen because all the relevant RCTs providing level 1 evidence on novel systemic therapies in addition to androgen deprivation in metastatic hormone-sensitive PCa have been published since 2015. Research terms that will be used for the research are the following: "(prostate cancer OR prostate adenocarcinoma) AND (metastatic)". Then, we will select only original articles reporting results from RCTs and we will exclude studies reporting results from the same RCT. We will collect the following patients' demographics and clinical characteristics:

- Age
- Race
- BMI
- Charlson comorbidity index
- Performance status (ECOG)
- Cardiovascular diseases
- Chronic kidney injury
- Liver diseases
- PSA
- Gleason score
- Extra-prostatic localization (pelvic lymph nodes, bone, others)

The published manuscript, supplementary files and ClinicalTrials.gov will be used to collect inclusion and exclusion criteria of each RCTs and other information such as:

- Drugs evaluated in the RCT
- Number of arms of the RCT
- Number of patients evaluated
- Number of patients included (overall and in each arm)
- Hospital setting (academic, non-academic)
- Country

### Target cohorts

Newly metastatic PCa will be identified in the predefined data sources as adult patients with newly diagnosed metastatic PCa with at least 365 days of prior observation. The following will be the inclusion criteria:

- Male adults (age $\geq$ 18)
- No other primary cancer except for non-melanoma skin cancer any time prior to index and within 30 days of index
- At least a diagnosis of metastatic disease (index date)
- At least one diagnosis of PCa dx with first PCa dx between 3 months prior and 1 month after INDEX (90 days prior or to 30 days to index)
- no history of PCa or prostate dysplasia within 365 days prior to index
- no drug exposure to ADT or androgen agonist/inhibitor within 365 days prior to index

### Features of interest

We will identify the following patients' demographics and clinical characteristics in the target cohort as assessed during the year (– 1 to – 365 days) pre-index date (PCa diagnosis):

- Age
- Race
- BMI
- Charlson comorbidity index
- Performance status (ECOG)
- Cardiovascular diseases
- Chronic kidney injury
- Liver diseases
- Bone marrow failure
- Adrenal insufficiency
- Diagnosis of HIV
- PSA
- Gleason score
- Extra-prostatic localization (lymph nodes, bone, central nervous system, others)

We will identify the treatment of the patients included in our cohort at or after their index date (0–365 days). This will be categorized as ADT alone, ADT plus radiotherapy, ADT plus androgen receptor targeted agents (ARTA), ADT plus chemotherapy, and ADT plus ARTA plus chemotherapy.

### Analysis: characterizing cohorts

All analyses will be performed using code developed for the OHDSI Methods library. A diagnostic package built off the OHDSI Cohort Diagnostics (https://ohdsi.github.io/CohortDiagnostics/) library, is included in the base package as a preliminary step to assess the fitness of use of phenotypes on your database. If a database passes cohort diagnostics, the full study package will be executed. Baseline covariates will be extracted using an optimized SQL extraction script based on principles of the Feature Extraction package (http://ohdsi.github.io/FeatureExtraction/) to quantify Demographics, Condition Group Eras, and Drug Group Eras Additional cohort-specific covariates will be constructed using OMOP Standard Vocabulary concepts. At the time of executing Feature Extraction, the package will create a data frame in which individuals' age will be extracted. Individuals' medical conditions, procedures, measurements and medications will be summarized (1) over the year prior to their index date (– 365 to 1 day), 2) at index date (0 day). Number and proportion of persons with feature variables during time-at-risk windows will be reported. Standardized mean differences (SMD) will be calculated when comparing characteristics of study cohorts, with plots comparing the mean values of characteristics for each of the characteristics (with the colour indicating the absolute value of the standardized difference of the mean).

Baseline disease characteristics at diagnosis will be reported using medians and proportions for non-normally distributed continuous variables and categorical variables, respectively.

We will compare the sample size of each study with the number of patients with newly diagnosed metastatic PCa identified in our cohort. Patient characteristics will be compared between patients included in the selected RCTs and those identified in our cohort using the Wilcoxon rank-sum and Fisher's exact test for continuous and categorical variables, respectively. Finally, we will assess the proportion of patients in our cohort who would have been eligible for RCT inclusion following the application of RCTs' inclusion/exclusion criteria.

Since available datasets may not represent the general population distribution because they may be affected by selection biases, additional analyses will be performed. Indeed, some of these databases are commercial claims databases containing data from insured individuals. Since the majority of these are active workers, the older part of the population could be underrepresented. Therefore, we will repeat our analysis in two different ways:

- Using subpopulations of randomly selected patients that match the US male population in terms of age distribution (median and interquartile range). Then, from this subpopulation, newly diagnosed metastatic PCa patients will be selected and we will compare their characteristics with those of patients included in the RCTs.
- Using a post-stratification weights method[24]. First, we stratify our dataset population by age (e.g. <45 years, 46–55 years, 56–65 years, 66–75 years, and > 75 years) and calculate the age group distribution. Using the male census data, we get the age group distribution in the US male population. We calculate the weights for each age group by dividing the proportion of that age group in the US population by proportion of that age group in our dataset. Then, we perform weighted analyses for the variables of interest.

### *Logistics of executing a federated analysis*

Sites will run the study analysis package locally on their data coded according to OMOP CDM. Only aggregate results will be shared with the study coordinator. Result files will be automatically staged into a ZIP file that can be transmitted using the OhdsiSharing R Library (http://ohdsi.github.io/OhdsiSharing/) or through a site's preferred SFTP client using a site-specific key provisioned by the OHDSI Study Coordinator. Local data stewards are encouraged to review study parameters to ensure minCellCount function follows local governance. At a minimum, it is encouraged to keep this value to greater than 5 to avoid any potential issues with re-identification of patients. (Note: covariates are constructed using controlled ontologies from the OMOP standard vocabularies though some labels may be replaced with publication-friendly labels due to space restrictions of the submitting journal).

### *Sample size and study power*

This study is undertaken using routinely collected data, all patients meeting the eligibility criteria above are included. No formal sample size and power calculation is performed.

## Strengths and limitations

### *Strengths*

The study will provide useful information on the generalizability of RCTs on newly diagnosed metastatic PCa using as a reference a large patient-level cohort of PCa patients, that should represent the real-life scenario. Indeed, data will be obtained from multiple centres and providers from at least five countries and two continents. Lastly, the use of routinely collected data from multiple sources maximizes the external validity and generalizability of the findings.

### *Limitations*

This study is carried out using data recorded in a collection of EHR, claims and tumour registries. As with any healthcare database used for secondary data analysis, the patient records might be incomplete in many respects and may have had erroneous entries, leading to misclassification of study variables. Data regarding diagnosis of PCa, treatments, pathology and lab results or baseline covariates prior to enrolment within the database may not be available. PCa specific characteristics such as stage or grade at diagnosis or the extent of the disease or mutational status of genes implicated in PCa are not readily available in some of the EHR and claims databases. Treatment provided in hospitals or any other setting outside each participating institution is not included. The available datasets may not represent the general population distribution because they may be affected by selection biases. Indeed, some of the available databases are commercial claims databases containing data form of insured individuals. Since the majority of these are active workers, the older part of the population could be underrepresented. For this reason, we will perform supplementary analyses as listed above. Medical conditions may be underestimated as they will be based on the presence of condition codes, with the absence of such a record taken to indicate the absence of a disease. Meanwhile, medication records indicate that an individual was prescribed or dispensed a particular drug, but this does not necessarily mean that an individual took the drug as originally prescribed or dispensed. A formal sample size and power analysis are not feasible due to the nature of the data and our hypothesis. Our data sources do not provide information regarding the hospital setting (e.g. academic vs. community hospital) where patients were treated. This prevents us from performing supplementary analyses on patients treated in a specific setting. Patient race is not commonly reported in our database, and therefore this characteristic cannot be compared between patients enroled in RCTs and those identified in our cohort. Finally, we cannot exclude that a minority of the patients included in the identified data sources might have been already included in RCTs. Therefore, they might have received therapies which were not considered as standard of care at the time of assessment.

## Protection of human subjects

The study uses only de-identified data. Confidentiality of patient records will be maintained at all times. Data custodians will remain in full control of executing the analysis and packaging results. There will be no transmission of patient-level data at any time during these analyses. Only aggregate statistics will be captured. Study packages will contain minimum cell count parameters to obscure any cells which fall below allowable reportable limits. All study reports will contain aggregate data only and will not identify individual patients or physicians.

## Management and reporting of adverse events and adverse reactions

According to the new guidelines for good pharmacovigilance practice (EMA/873138/2011)[25] and ISPE[26], there is no requirement for expedited reporting of adverse drug reactions from studies with secondary use of data (such as electronic healthcare databases).

## Plans for disseminating and communicating study results

The results of the study will be presented at international Urological meetings in the form of abstracts. The final results will be published as full-text paper in an international peer-reviewed urological journal.

Results of this study will be published following guidelines, including those for authorship, established by the International Committee of Medical Journal Editors[27]. When reporting results of this study, the appropriate Strengthening the Reporting of Observational Studies in Epidemiology checklist will be followed[27].

## Ethical approval

Not applicable.

## Consent

Not applicable.

## Source of funding

The study is supported by the Innovative Medicines Initiative 2 Joint Undertaking projects PIONEER (grant agreement No 777492) and EHDEN (grant agreement No 806968). IMI2 receives support from the European Union's Horizon 2020 research program and the European Federation of Pharmaceutical Industries and Associations (EFPIA). Many of the contributors are part of OHDSI, a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics and may have other funding sources which will be listed in the study manuscripts.

## Author contribution

Conceptualization: G.G., F.P., J.N'D. Formal analysis: B.D.M., A.G., A.H. Writing—original draft preparation: G.G, F.P. Supervision: A.B, J.N'D, J.G.-R., T.A., R.N. Project administration: C.S., S.E.-A. All authors have read and agreed to the published version of the manuscript.

## Conflicts of interest disclosure

No.

## Research registration unique identifying number (UIN)

Not applicable.

## Guarantor

Giorgio Gandaglia.

## Data availability statement

In the interest of transparency and scientific reproducibility, all study materials including the computer-executable code (which is compatible with any dataset in the OMOP common data model) will be made available.

## Provenance and peer review

No.

## References

[1] Rawla P. Epidemiology of prostate cancer. World J Oncol 2019;10:63–89.
[2] Gandaglia G, Leni R, Bray F, *et al*. Epidemiology and prevention of prostate cancer. Eur Urol Oncol 2021;4:877–92.
[3] Dasgupta P, Baade PD, Aitken JF, *et al*. Geographical variations in prostate cancer outcomes: a systematic review of international evidence. Front Oncol 2019;9:238.
[4] Scher HI, Solo K, Valant J, *et al*. Prevalence of prostate cancer clinical states and mortality in the United States: estimates using a dynamic progression model. PLoS One 2015;10:e0139440.
[5] Cornford P, van den Bergh RCN, Briers E, *et al*. EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer. Part II—2020 Update: Treatment of Relapsing and Metastatic Prostate Cancer. Eur Urol 2021;79:263–82.
[6] Lowrance W, Dreicer R, Jarrard DF, *et al*. Updates to Advanced Prostate Cancer: AUA/SUO Guideline (2023). J Urol 2023;209:1082–90.
[7] Armstrong AJ, Szmulewitz RZ, Petrylak DP, *et al*. ARCHES: A Randomized, Phase III Study of Androgen Deprivation Therapy With Enzalutamide or Placebo in Men With Metastatic Hormone-Sensitive Prostate Cancer. J Clin Oncol 2019;37:2974–86.
[8] Sweeney CJ, Chen Y-H, Carducci M, *et al*. Chemohormonal therapy in metastatic hormone-sensitive prostate cancer. N Engl J Med 2015;373:737–46.
[9] Davis ID, Martin AJ, Stockler MR, *et al*. Enzalutamide with standard first-line therapy in metastatic prostate cancer. N Engl J Med 2019;381:121–31.
[10] Gravis G, Fizazi K, Joly F, *et al*. Androgen-deprivation therapy alone or with docetaxel in non-castrate metastatic prostate cancer (GETUG-AFU 15): a randomised, open-label, phase 3 trial. Lancet Oncol 2013;14:149–58.
[11] Boevé LMS, Hulshof MCCM, Vis AN, *et al*. Effect on survival of androgen deprivation therapy alone compared to androgen deprivation therapy combined with concurrent radiation therapy to the prostate in patients with primary bone metastatic prostate cancer in a prospective randomised clinical trial: data from the HORRAD Trial. Eur Urol 2019;75:410–8.
[12] Fizazi K, Tran N, Fein L, *et al*. Abiraterone plus prednisone in metastatic, castration-sensitive prostate cancer. N Engl J Med 2017;377:352–60.
[13] Agarwal N, Tangen CM, Hussain MHA, *et al*. Orteronel for metastatic hormone-sensitive prostate cancer: a multicenter, randomized, open-label phase III Trial (SWOG-1216). J Clin Oncol 2022;40:3301–9.
[14] Vaishampayan UN, Heilbrun LK, Monk P, *et al*. Clinical efficacy of enzalutamide vs bicalutamide combined with androgen deprivation therapy in men with metastatic hormone-sensitive prostate cancer: a randomized clinical trial. JAMA Netw Open 2021;4:e2034633.
[15] Fizazi K, Foulon S, Carles J, *et al*. Abiraterone plus prednisone added to androgen deprivation therapy and docetaxel in de novo metastatic castration-sensitive prostate cancer (PEACE-1): a multicentre, open-label, randomised, phase 3 study with a 2 × 2 factorial design. Lancet 2022;399:1695–707.
[16] James ND, de Bono JS, Spears MR, *et al*. Abiraterone for prostate cancer not previously treated with hormone therapy. N Engl J Med 2017;377:338–51.
[17] Chi KN, Agarwal N, Bjartell A, *et al*. Apalutamide for metastatic, castration-sensitive prostate cancer. N Engl J Med 2019;381:661.
[18] Smith MR, Hussain M, Saad F, *et al*. Darolutamide and survival in metastatic, hormone-sensitive prostate cancer. N Engl J Med 2022;386:1132–42.
[19] Guyatt G, Oxman AD, Akl EA, *et al*. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383–94.
[20] Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply? Lancet 2005;365:82–93.
[21] Kennedy-Martin T, Curtis S, Faries D, *et al*. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials 2015;16:1–14.
[22] Monti S, Grosso V, Todoerti M, *et al*. Randomized controlled trials and real-world data: differences and similarities to untangle literature data. Rheumatology (Oxford) 2018;57:vii54–8.

[23] Omar MI, Roobol MJ, Ribal MJ, *et al*. Introducing PIONEER: a project to harness big data in prostate cancer research. Nat Rev Urol 2020;17:6; 2020;17:351–362.
[24] Downes M, Gurrin LC, English DR, *et al*. Multilevel regression and post-stratification: a modeling approach to estimating population quantities from highly selected survey samples. Am J Epidemiol 2018;187:1780–90.
[25] European Medicines Agency. Guidelines on good pharmacovigilance practices (GVP): Module VIII – Post-authorisation safety studies. 2017. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-gvp-module-viii-post-authorisation-safety-studies-rev-3_en.pdf
[26] Hallas J. Guidelines for good pharmacoepidemiology practice (GPP). Pharmacoepidemiol Drug Saf 2016;25:2–10.
[27] ICMJE. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. 2019. https://www.icmje.org/icmje-recommendations.pdf