JCO Clinical Cancer Informatics May 2019 No 3 1-12
https://doi.org/10.1200/CCI.18.00144

PROSPeCT: A Predictive Research Online System for Prostate Cancer Tasks

Maria Cutumisu, PhD; Catalina Vasquez, MSc; Maxwell Uhlich; Perrin H. Beatty, PhD;
Homeira Hamayeli-Mehrabani, PhD; Rume Djebah, MBBS, MHA;
Albert Murtha, MD, FRCPC; Russell Greiner, PhD; John D. Lewis*, PhD

University of Alberta, Canada

* Corresponding author: John D. Lewis, PhD, Department of Experimental Oncology, University of Alberta, Edmonton, AB T6G 2E1 Canada; e-mail: jdlewis@ualberta.ca

CONTEXT

- **Key Objective:** We introduce and describe the Predictive Research Online System Prostate Cancer Tasks (PROSPeCT), an online clinical information system.
- **Knowledge Generated:** PROSPeCT provides a visual interface aiding researchers and clinicians in generating hypotheses and constructing prostate cancer-related queries, identifies patient cohorts with specified characteristics; aids clinicians in managing individual patients by providing possible diagnosis, treatment plans, outcomes, and identifying possible complications; and interrogates patient information from the Alberta Prostate Cancer Registry hosted by the Alberta Prostate Cancer Research Initiative. We also compare PROSPeCT with related digital tools.
- **Relevance:** The PROSPeCT system can be applied in a clinical context to facilitate the work of researchers and clinicians in managing and improving prostate cancer outcomes for patients.

Abstract

Purpose An online clinical information system, called Predictive Research Online System Prostate Cancer Tasks (PROSPeCT), was developed to enable users to query the Alberta Prostate Cancer Registry database hosted by the Alberta Prostate Cancer Research Initiative. To deliver high-quality patient treatment, prostate cancer clinicians and researchers require a user-friendly system that offers an easy and efficient way to obtain relevant and accurate information about patients from a robust and expanding database.

Methods PROSPeCT was designed and implemented to make it easy for users to query the prostate cancer patient database by creating, saving, and reusing simple and complex definitions. We describe its intuitive nature by exemplifying the creation and use of a complex definition to identify a "high-risk" patient cohort.

Results PROSPeCT was made to minimize user error and to maximize efficiency without requiring the user to have programming skills. Thus, it provides tools that allow both novice and expert users to easily identify patient cohorts, manage individual patient care, perform Kaplan-Meier estimates, plot aggregate PSA views, compute PSA-doubling time, and visualize results.

Conclusion This report provides an overview of PROSPeCT, a system that helps clinicians to identify appropriate patient treatments and researchers to develop prostate cancer hypotheses, with the overarching goal of improving the quality of life of patients with prostate cancer. We have made available the code for the PROSPeCT implementation at https://github.com/max-uhlich/e-PROSPeCT.

Keywords. APCaRI, Alberta Prostate Cancer Registry, Clinical data, Clinical Information System, Kaplan-Meier estimator, prostate cancer, PROSPeCT, PSA doubling time, PSA trajectory, REDCap

INTRODUCTION

Globally, prostate cancer (PCa) is the second most commonly diagnosed cancer. In the developed world, it is the most common cancer in men,[1] with similar statistics for Canada[2] and the United States.[3] Although the PCa survival rate has increased overall, the incidence of PCa is rising, triggering increased cost of treatment[4] with significant repercussions to the global economy. PCa is being detected earlier due to the use of the prostate-specific antigen (PSA) blood test, which, unfortunately, leads to overdiagnosis and excessive use of invasive treatment methods, even in low-risk patients with cancer. The most influential driver of patient survivability is the quality of health care, including the management of the disease,[5-7] informed by the results of research on patient cohorts. For clinicians, the use of an active surveillance strategy with low-risk patients has been proposed as an alternative to invasive, unnecessary therapies.[8] Being able to accurately distinguish low-risk from high-risk patients with PCa is critical for this approach.

Consequently, it is important to develop tools that help clinicians to effectively treat patients and help medical researchers to identify relevant patient cohorts and to answer questions about such groups. There are different types of digital query-based tools that clinicians and researchers can use to access patient information, clinical standards, best healthcare practices, literature reviews, and to analyze patient or cohort data for diagnostic or hypothesis-building purposes. These range from general spreadsheet-based programs to relational database management systems to highly specialized algorithm and clinical information systems (CISs),[9-12] where the user either constructs the query logic manually to perform database queries or uses query-building tools.[13]

For clinicians, the tools must support the collection and management of up-to-date patient data, because patients with PCa generally require regular testing after initial diagnosis to estimate individual risk. These tools are used to inform decisions regarding the appropriate course of treatment. For medical researchers, such tools should help identify and answer relevant research questions on the basis of regularly updated data. Clinicians and researchers need a query-building platform that is intuitive to use and does not require knowledge of query languages or of database models.[10,11,13] Currently, there are many general-purpose tools that allow clinicians and researchers to query databases of patient with PCa, such as cancer information systems; these tools vary greatly in terms of ease of use and analytical power.[14] The purpose of this report is to describe one such tool called Predictive Research Online System Prostate Cancer Tasks (PROSPeCT).[15]

Here, we introduce and describe PROSPeCT, an online CIS that (1) provides a visual interface aiding researchers and clinicians in generating hypotheses and constructing PCa-related queries; (2) identifies patient cohorts with specified characteristics; (3) aids clinicians in managing a single patient by providing possible diagnosis, treatment plans, outcomes, and identifying possible complications; and (4) interrogates patient information from the Alberta Prostate Cancer Research Initiative (APCaRI).[16] And we compare PROSPeCT with related digital tools.

METHODS

PROSPeCT is a web application powered by the Apache Tomcat web server. It is implemented using Java Google Web Toolkit (http://www.gwtproject.org/) modules and it integrates the Alberta Prostate Cancer Registry PostgreSQL database[17] and APCaRI Python import modules.[18] Many of its data visualization components also use the D3.js JavaScript library, as detailed in the *Data Supplement*.

Rationale for Creating the User-Friendly, Query-Building Tool: PROSPeCT

PROSPeCT was designed to meet the needs and preferences of the target users: clinicians and researchers. The main features driving its creation are outlined as follows.

1. **Query Facilitation**. Relevant subsets of patients should be easy to identify from the database. This may require complex queries but needs to be feasible without the use of complex text-based programming.
2. **Incorporation of Predefined Features**. Complicated characteristics or features, such as PCa risk stratification and PCa recurrence or progression, should be predefined and, therefore, easy to incorporate into the queries. These features often require defining terms that correspond to a specific time interval for the query. Thus, the system must include predefined features (e.g., Interval Query) to allow the user to easily define and use time intervals (e.g., from birth to onset).
3. **Cohort Visualization**. Important results about the identified patient subpopulations should be easily visualized with general dashboard summaries that display useful population characteristics from large data sets in a scalable way. It must also be easy to prepare commonly-used graphical images (e.g., Kaplan-Meier curves, aggregated PSA views).
4. **Individual Visualization**. Important results about individual patients should be easily visualized with the “Single Patient Summary” that displays useful patient characteristics, including commonly used graphical images, such as PSA plots and the PSA doubling time (PSADT) between pairs of time points.

RESULTS

Dataflow

PROSPeCT allows users to query the database of patient information. The dataflow to produce and maintain that database is a stepwise process originating with the patient and progressing through the clinic, to the Research Electronic Data Capture (REDCap; https://www.project-redcap.org)-managed Alberta Prostate Cancer Registry, and to PROSPeCT, which allows querying by the clinicians and researchers (Fig. 1).

Once the patient provides consent to be part of the study, REDCap captures (1) patient-reported information, (2) PCa diagnosis-related data, (3) disease-specific information, (4) results from biomarker analysis, and (5) inventory of biosamples collected for the Alberta Prostate Cancer Registry and Biorepository. Note that some features have multiple time-linked values. This information includes data fields, such as demographic, comorbidities, use of medications, clinical, patient-reported outcomes, and biosample information. The data in REDCap are downloaded, processed, and uploaded once per week into PROSPeCT using five APCaRI Python modules.

Overview of the PROSPeCT Interface

Once the data are imported, users can probe it through the interface that contains five panes, shown in Fig. 2: Fields (primary and PROSPeCT-derived); Definition Builder; Defined Populations; Operations; and Results.

Fields

Primary fields contain unprocessed Alberta Prostate Cancer Registry patient data that are imported directly from the REDCap-managed database, which includes patient demographics, medical history, clinical results including prostate specific antigen (PSA) levels, biopsy results, treatment information and outcomes, biosample information, and quality-of-life surveys.

1. Incorporation of Predefined Features

PROSPeCT-derived fields are computed offline by the import modules during the data transfer from REDCap and include PCa recurrence or progression (defined here by biochemical recurrence or PSA failure) and risk stratification. PCa recurrence/progression in men treated with prostatectomy or radiation therapy is computed based on their PSA levels over time, graphed as a PSA trajectory plot and visualized in the Single Patient Summary.

PROSPeCT also provides several browser-computed (i.e., on-the-fly) data points, such as the patient's age as computed from their birth date and today's date, and the PSA doubling time (PSADT) in the PSA trajectory graph from the Single Patient Summary.

2. Query Facilitation: Definitions

One of the greatest strengths of PROSPeCT is the set of tools it provides to allow users to easily create, store, and re-use their own definitions, by sequentially selecting a field (from the Fields Pane; see Fig. 2A) and dragging and dropping that field into the Definitions Builder Pane. For each field, the user then chooses an operator (e.g., =, <>, <, <=, >, >=, IS NULL, IS NOT NULL) and a value. Definitions are listed in the Definitions Builder Pane and they can be simple, using one field, such as "Clinical staging (T) = T1" or complex, employing two or more fields, which involve arbitrary Boolean combinations.

Fig. 3 depicts a complex user-created definition based on four fields, designed to identify the patient cohort: "All patients whose total Gleason score (overall) is at least 8, or whose PSA level is over 20, or whose clinical staging is either T3 or T4". First, the user dragged and dropped the "Total Gleason score (overall)" field from the Fields Pane to the Definition Builder Pane and chose the mathematical operator “>=” from a drop-down menu, entering the value ”8” in the adjacent field. Second, the user dragged and dropped the PSA field from the Fields Pane to the Definition Builder Pane and again chose the correct operator, entering the value associated with the PSA level (i.e., “> 20”). The user continued building the definition with the “Clinical staging” field, choosing the necessary operators, and inputting the required values. Because the user-created definition in Fig. 3 contained four fields combined by “or” instead of “and”, the user must choose that operator in the Definition Builder Pane. When the user then clicked on “Create Definition”, PROSPeCT automatically identified those instances from the data set and displayed them in a table in the Results Pane. This definition is also summarized in the “Defined Populations Pane,” where the user is able to provide a definition name, such as “High_Risk.” If the definition is needed for later use, the user can right click on the blue label and click 'Save this Definition' in the drop-down menu that appears.

3. Cohort Visualization

Once the user has queried the database and identified an individual patient or a patient cohort, the user can visualize the results in the Results Pane, which contains tabs of types: Table, Kaplan-Meier Estimator, Aggregate PSA View, Single Patient Summary, or Cohort/Group Statistics Dashboard (Fig. 4). The user can create a new tab by clicking the relevant button in the Operations Pane.

The bottom of Fig. 2 shows the “Table” type (in the Results part of the display), which displays the result of performing the union of all the Defined Populations selected in the Defined Population Pane. The rows include all the instances selected by any of the definitions; the columns include all the features mentioned by any of these definitions. Notably, PROSPeCT also allows the user to easily add another column (field) to the current list of patients by selecting the new field from the Fields Pane and dragging and dropping it onto the Results (Table) pane to populate that column with the values of the associated feature, for each patient listed.

The Kaplan-Meier Estimator and the Aggregate PSA View are built-in applications in PROSPeCT that greatly aid PCa treatment decisions (Fig. 4A and 4B)[19]. The Aggregate PSA View produces a graph superimposing all PSA trajectories of the current defined population-patient cohort over time, centered at a user-selected date (e.g., date of biopsy; Fig. 4B).

4. Individual Visualization

The Single Patient Summary allows the user to generate a page that summarizes important aspects associated with a patient (Fig. 4D). This summary includes a graph of the PSA trajectory for the patient over time (Fig. 4C). Fig. 4D superimposes five events over that plot, and Fig. 4E shows that the user can easily compute the PSADT (e.g., “32 months” between January 2015 and June 2015). The Cohort/Group Statistics Dashboard application generates graphs comparing the data from fields across patient cohorts (Fig. 4C).

Thus, PROSPeCT offers a convenient, easy, and fast method of querying the database to identify and examine the specified patient cohorts. It was designed to be used with very little training, due to the drag-and-drop interface and the preprogramming of many complex PCa-specific fields for the user.

Comparison of PROSPeCT with Other Digital Tools

There are many web-based tools that examine the generalized populations of patients with cancer (i.e., general-population statistics) based on the SEER,[20] National Cancer Database,[21] or the University of California, San Francisco Cancer of the Prostate Strategic Urologic Research Endeaver (CaPSURE)[22] databases. In contrast, PROSPeCT examines individuals within a specified cohort. Similar to PROSPeCT, even though other databases can be queried by many web-based tools that also offer survival reports, calculate survival by stage at diagnosis, and determine trends and incidence rates for various cancer sites over time, some databases store only PCa data (e.g., CaPSURE), whereas others include data for other types of cancer as well (e.g., SEER, National Cancer Database).

Although desktop-based programs such as Microsoft Excel or Microsoft Access (Redmond, CA) can help analyze patient data extracted from a small-scale database, this analysis can also take more computer time per query, be prone to user errors, restrict the input variables to predefined formats, and impose a flat data model that cannot easily capture relationships detectable by complex queries.[9,13] Automated extraction of patient cohorts from well-populated databases through user-defined queries would improve the accuracy of data retrieval, which, in turn, would greatly reduce the time spent by clinicians and researchers on data collection and analysis.

Our overarching goal is to create a general-purpose system that can help answer a range of questions. This differs from special-purpose systems that can answer only a single, pre-defined question. Gregg et al.[23] recently developed a natural-language processing algorithm to extract PCa risk stratification data from existing electronic medical record databases. This allows clinicians to characterize PCa disease risk. Researchers hypothesized that PCa risk groups could be accurately determined from natural-language processing-extracted data in at least 90% of patients. Although extremely useful, this tool could only generate results for this specific question.

Table 1 compares PROSPeCT with several other digital query tools on the basis of ease of use, generation and visualization of results, data management, security, and cost. The tools can store a data set and perform queries on that data set, but differ in scale, ease of use, and flexibility in their programmed capabilities. Clinicians and researchers can greatly benefit from using PROSPeCT to query the Alberta Prostate Cancer Registry (or any other compatible PCa database - because it was designed with a user-friendly interface, it can generate query results quickly, and it can allow users to perform PCa-specific data calculations and visualizations via built-in applications.

We also compared the time spent by a user to query the Alberta Prostate Cancer Registry database with a complex, multifield definition using REDCap, Microsoft Excel, and PROSPeCT, asking the same user to run the same complex query with the three tools: "interrogate the whole population to extract features on specific patient cohorts with PCa disease-recurrence after receiving specific treatments." REDCap was not able to complete the query, because it is designed for retrieving records by field or table and it did not have any option to use multiple dates from different timelines and query forms. Microsoft Excel was able to perform this query, but the user had to write and implement extensive macro programming to generate the defined population list. Because PROSPeCT contains prederived PSA recurrence fields, the user could quickly build the definition used for this query and generate the defined population that fits this query, spending much less time using PROSPeCT to build this definition and produce the query report, compared to using Excel (Table 1).

DISCUSSION

The PROSPeCT interface is designed to make it easy for clinicians and researchers to interrogate the large Alberta Prostate Cancer Registry database of more than 3600 individuals, with more than 1500 features, and to save relevant definitions for use on updated versions of this database. There are many advantages of using PROSPeCT as an online database interface relative to other digital tools. First, it uses a drag-and-drop paradigm, so it is easy and quick for users with no programming expertise to build complex definitions to query the database. In contrast, as Microsoft Excel requires the user to write macros to run complex queries, Excel users are likely to spend much more time generating queries, in comparison to PROSPeCT users. Second, it contains several precomputed fields relevant to Pca (e.g., PCa risk stratification and PCa recurrence or progression) that allow users to quickly identify relevant cohorts of patients and then important characteristics of these patients. Third, it contains several built-in applications relevant to PCa, such as the Kaplan-Meier estimator, Aggregated PSA View, PSADT, and Interval Query. Even though the PSADT is relatively simple to calculate, and clinicians and researchers find that the doubling-time information is useful for comparing treatment options and research studies, it is difficult to compute in Excel. This is due to the way data is organized in spreadsheets, as it requires users to extract multiple patient-specific data points from the database, then compute the doubling time for each chosen pair of data points. This complexity is why many clinicians do not use it to compute PSADT.[3] However, the PSADT facility incorporated in "Single Patient Summary" makes this extrapolated result easily obtainable when using the PROSPeCT interface. Finally, it allows users to easily visualize single patient summaries, including a plot of the patient's PSA values over time with overlays showing relevant events, and to compare defined populations using the cohort/group dashboard feature.

Limitations. First, the current system is limited by the synchronization with the REDCap data management tool, which is not automated, because REDCap does not provide the facilities needed to make this process seamless. Second, although, in general, a user does not require computer programming skills to use PROSPeCT, some tasks need to be defined by a programmer (e.g., PSA Failure). Also, although the PROSPeCT system does have an adjustable template for its dashboard (Fig. S1), this action currently requires a programmer to adjust. Third, although PROSPeCT was designed to maximize expressiveness while minimizing complexity, there is a boundary where certain queries are not expressible by the system. This stems from PROSPeCT trying to strike a balance between complex and expressive query languages like Standard Query Language (SQL), which require much training to use, and simpler, visual interfaces that can be used with very little training. Fourth, this research is limited by the nature of the data stored in the PCa database, because the queries depend on definitions that draw on basic database fields. Finally, although informal tests of accuracy and efficiency have been conducted, this report focuses on descriptive and not empirical research.

Future Work. The current PROSPeCT system gets patient values on a scheduled basis. The near-term goal for PROSPeCT is to have a direct, real-time link to electronic medical records, meaning it could be used to generate high-value data analysis to aid with treatment decision-making for a current patient. The long-term goal involves connecting PROSPeCT with databases for other types of cancer or diseases, to allow cross-referencing of patient data that could dramatically increase the effectiveness of PROSPeCT as both a diagnostic and research tool for PCa and other diseases. Overall, PROSPeCT enables users to easily and quickly create elaborate, error-free queries, especially those related to PCa.

Acknowledgements. We thank Nawaid Usmani, MD, for expert guidance with clinical features.

Disclaimers. The views expressed in the submitted article are those of the authors and not an official position of the institution or funder.

Support. Supported by Alberta Cancer Foundation Program Grant No. 26491. J.D.L. holds the Frank and Carla Sojonky Chair in Prostate Cancer Research funded by the Alberta Cancer Foundation.

Figure legends

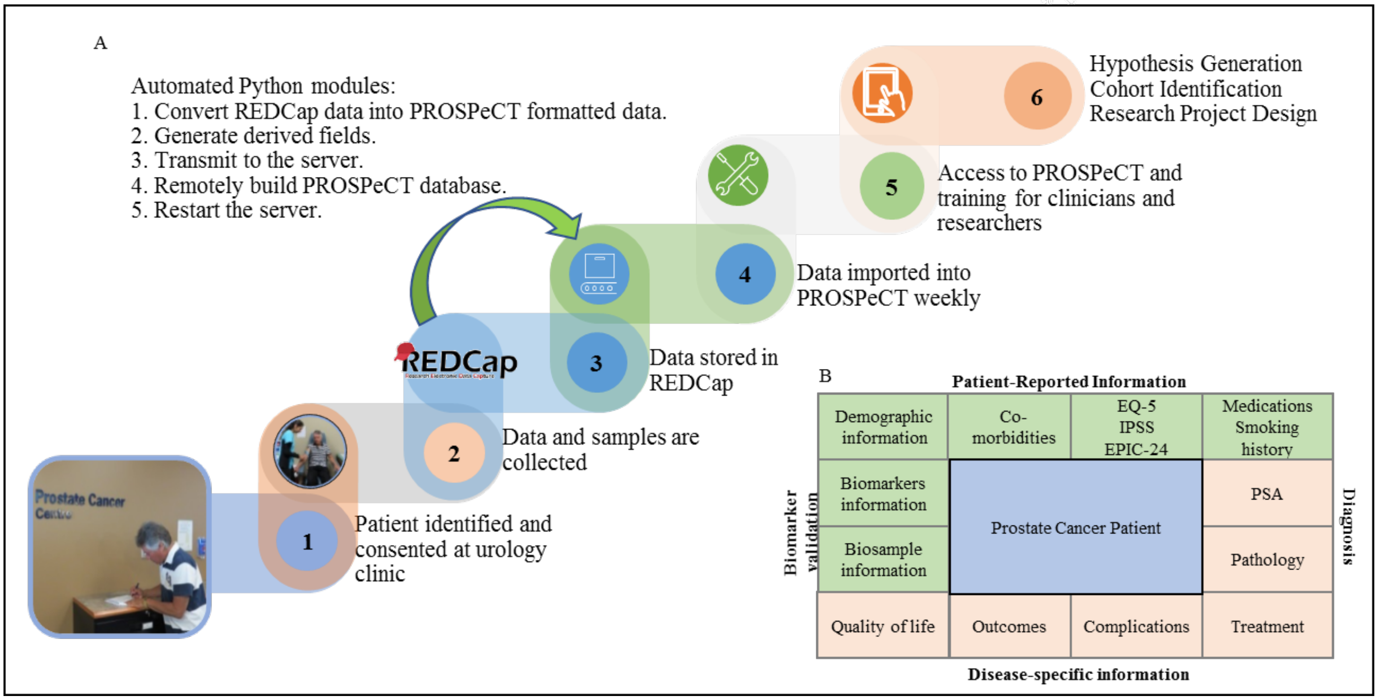
Fig. 1. The data flow from Research Electronic Data Capture (REDCap) to Predictive Research Online System Prostate Cancer Tasks (PROSPeCT) and the types of data and analysis available with PROSPeCT. (A) The data file is manually exported from REDCap, then processed by automated modules that extract fields and update the database, and finally imported into the PROSPeCT web application. (B) The types of reported patient information and analysis available in PROSPeCT. EPIC-26, Expanded Prostate Cancer Index Composite; EQ-5D, EuroQol Group Standardized Instrument; IPSS, International Prostate Symptom Score; PSA, prostate-specific antigen.

Fig. 2. The Predictive Research Online System Prostate Cancer Tasks (PROSPeCT) interface used to query the database and visualize the results. Users can see the following panes on their monitors simultaneously: (A) Fields; (B) Definitions Builder; (C) Defined Populations; (D) Operations; and (E) Results.

Fig. 3. The use of the Predictive Research Online System Prostate Cancer Tasks (PROSPeCT) interface is illustrated by outlining the steps required to create the multifield definition called “High-Risk” into the Definition Builder Pane, using fields in the Fields Pane. The definition is then listed in the Defined Populations Pane. The query based on this definition generates the list

of patients satisfying the High_Risk definition (the Defined Population) and displays their Gleason scores, PSA values, and clinical staging values in the Results Pane. The Fields Pane includes the primary and Research Electronic Data Capture (REDCap)-derived fields as well as the Derived Fields computed by PROSPeCT. The tab called “Guest’s Saved Definitions” contains the user-created definitions stored by the user for future use. PSA, prostate-specific antigen.

Fig. 4. The Operations Pane with the built-in application options for listing, exporting, and analyzing query results from definitions and summarizing single patient data. (A) Kaplan-Meier Estimator-generated Kaplan-Meier curve from a high-risk and a low-risk patient cohort. (B) PSA aggregate trajectory plot of eight patients. (C) Patient cohort summary dashboard. (D) Single patient summary with a PSA trajectory plot. (E) A single patient PSA trajectory plot showing the calculated PSA doubling time. PSA, prostate-specific antigen.



Automated Python modules:
1. Convert REDCap data into PROSPeCT formatted data.
2. Generate derived fields.
3. Transmit to the server.
4. Remotely build PROSPeCT database.
5. Restart the server.

1 Patient identified and consented at urology clinic

2 Data and samples are collected

3 Data stored in REDCap

4 Data imported into PROSPeCT weekly

5 Access to PROSPeCT and training for clinicians and researchers

6 Hypothesis Generation
Cohort Identification
Research Project Design

Patient-Reported Information

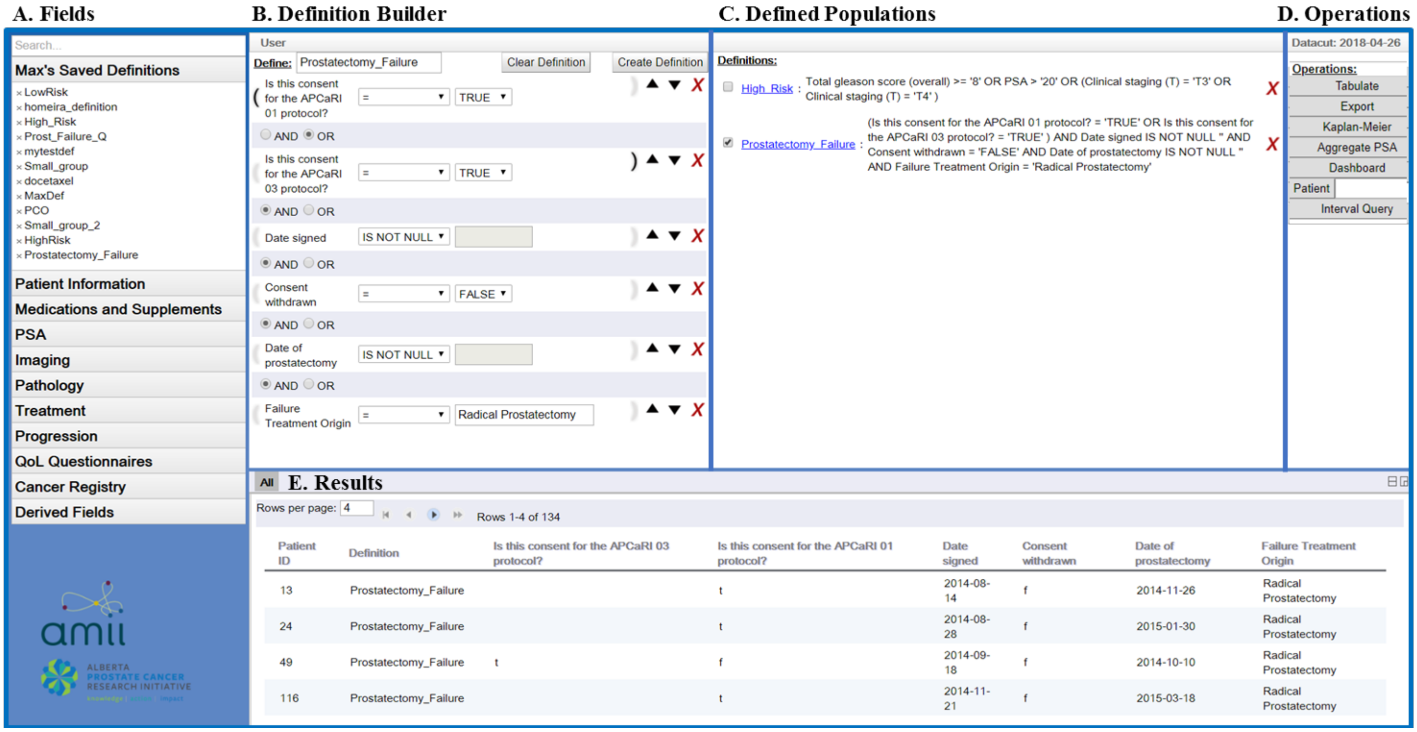
Biomarker validation

Diagnosis

Demographic information	Co-morbidities	EQ-5 IPSS EPIC-24	Medications Smoking history
Biomarkers information	Prostate Cancer Patient		PSA
Biosample information			Pathology
Quality of life	Outcomes	Complications	Treatment

Disease-specific information

Fig. 1.



Search...

Max's Saved Definitions

- × LowRisk
- × homeira_definition
- × High_Risk
- × Prost_Failure_Q
- × mytestdef
- × Small_group
- × docetaxel
- × MaxDef
- × PCO
- × Small_group_2
- × HighRisk
- × Prostatectomy_Failure

Patient Information
Medications and Supplements
PSA
Imaging
Pathology
Treatment
Progression
QoL Questionnaires
Cancer Registry
Derived Fields

User

Define: Prostatectomy_Failure Clear Definition Create Definition

(Is this consent for the APCaRI 01 protocol? = TRUE
○ AND ● OR
Is this consent for the APCaRI 03 protocol? = TRUE)
● AND ○ OR
Date signed IS NOT NULL
● AND ○ OR
Consent withdrawn = FALSE
● AND ○ OR
Date of prostatectomy IS NOT NULL
● AND ○ OR
Failure Treatment Origin = Radical Prostatectomy

Definitions:

□ High_Risk : Total gleason score (overall) >= '8' OR PSA > '20' OR (Clinical staging (T) = 'T3' OR Clinical staging (T) = 'T4')

☑ Prostatectomy_Failure : (Is this consent for the APCaRI 01 protocol? = 'TRUE' OR Is this consent for the APCaRI 03 protocol? = 'TRUE') AND Date signed IS NOT NULL '' AND Consent withdrawn = 'FALSE' AND Date of prostatectomy IS NOT NULL '' AND Failure Treatment Origin = 'Radical Prostatectomy'

Datacut: 2018-04-26

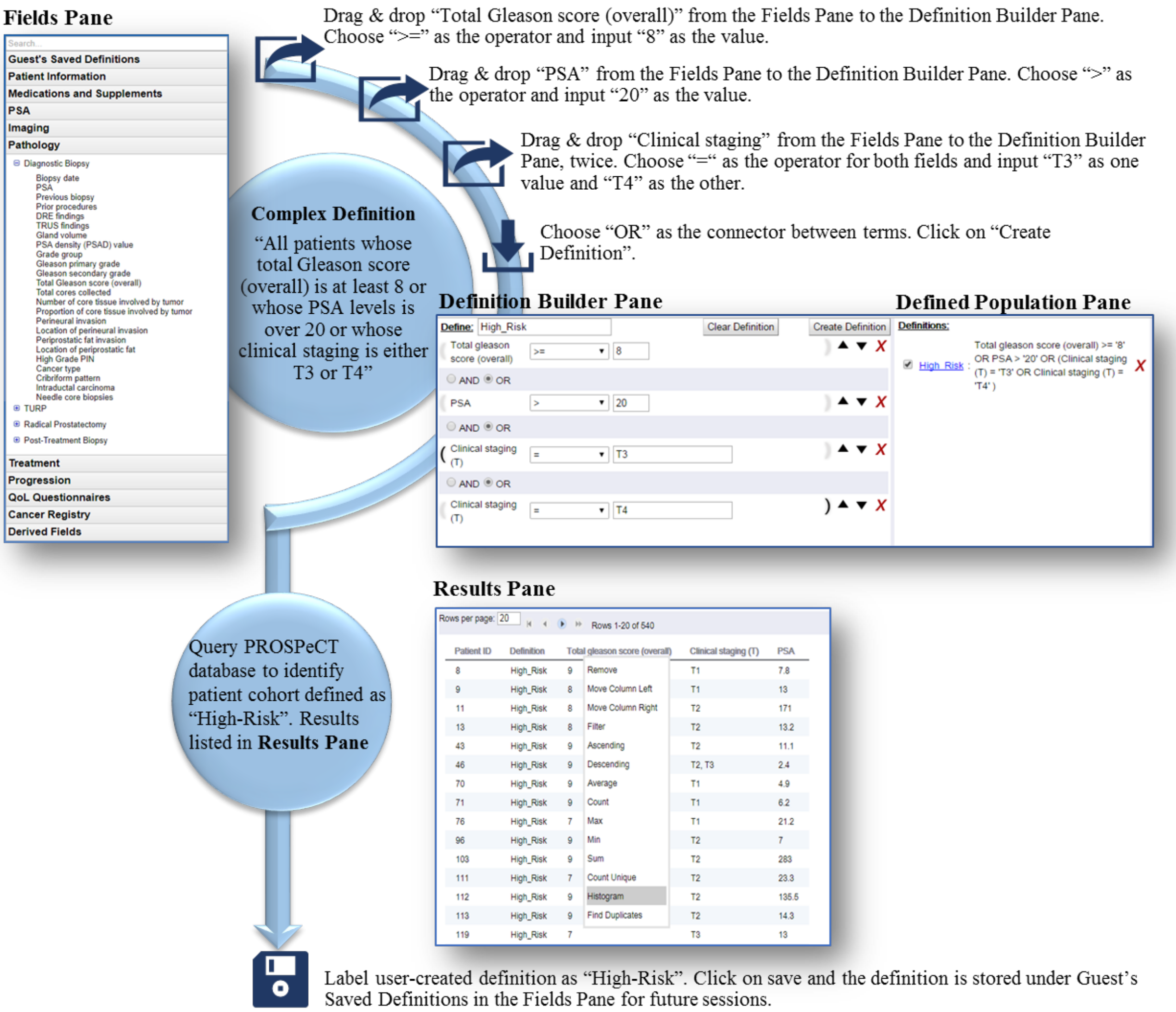
Operations:
Tabulate
Export
Kaplan-Meier
Aggregate PSA
Dashboard
Patient
Interval Query

All

Rows per page: 4 Rows 1-4 of 134

Patient ID	Definition	Is this consent for the APCaRI 03 protocol?	Is this consent for the APCaRI 01 protocol?	Date signed	Consent withdrawn	Date of prostatectomy	Failure Treatment Origin
13	Prostatectomy_Failure		t	2014-08-14	f	2014-11-26	Radical Prostatectomy
24	Prostatectomy_Failure		t	2014-08-28	f	2015-01-30	Radical Prostatectomy
49	Prostatectomy_Failure	t	f	2014-09-18	f	2014-10-10	Radical Prostatectomy
116	Prostatectomy_Failure		t	2014-11-21	f	2015-03-18	Radical Prostatectomy

Fig. 2.



Drag & drop "Total Gleason score (overall)" from the Fields Pane to the Definition Builder Pane. Choose ">=" as the operator and input "8" as the value.

Drag & drop "PSA" from the Fields Pane to the Definition Builder Pane. Choose ">" as the operator and input "20" as the value.

Drag & drop "Clinical staging" from the Fields Pane to the Definition Builder Pane, twice. Choose "=" as the operator for both fields and input "T3" as one value and "T4" as the other.

Complex Definition

"All patients whose total Gleason score (overall) is at least 8 or whose PSA levels is over 20 or whose clinical staging is either T3 or T4"

Choose "OR" as the connector between terms. Click on "Create Definition".

Definition Builder Pane

Define: High_Risk

(Total gleason score (overall) >= 8) ▲ ▼ X

○ AND ● OR

(PSA > 20) ▲ ▼ X

○ AND ● OR

(Clinical staging (T) = T3) ▲ ▼ X

○ AND ● OR

(Clinical staging (T) = T4) ▲ ▼ X

Defined Population Pane

Definitions:

☑ High_Risk : Total gleason score (overall) >= '8' OR PSA > '20' OR (Clinical staging (T) = 'T3' OR Clinical staging (T) = 'T4') X

Query PROSPeCT database to identify patient cohort defined as "High-Risk". Results listed in **Results Pane**

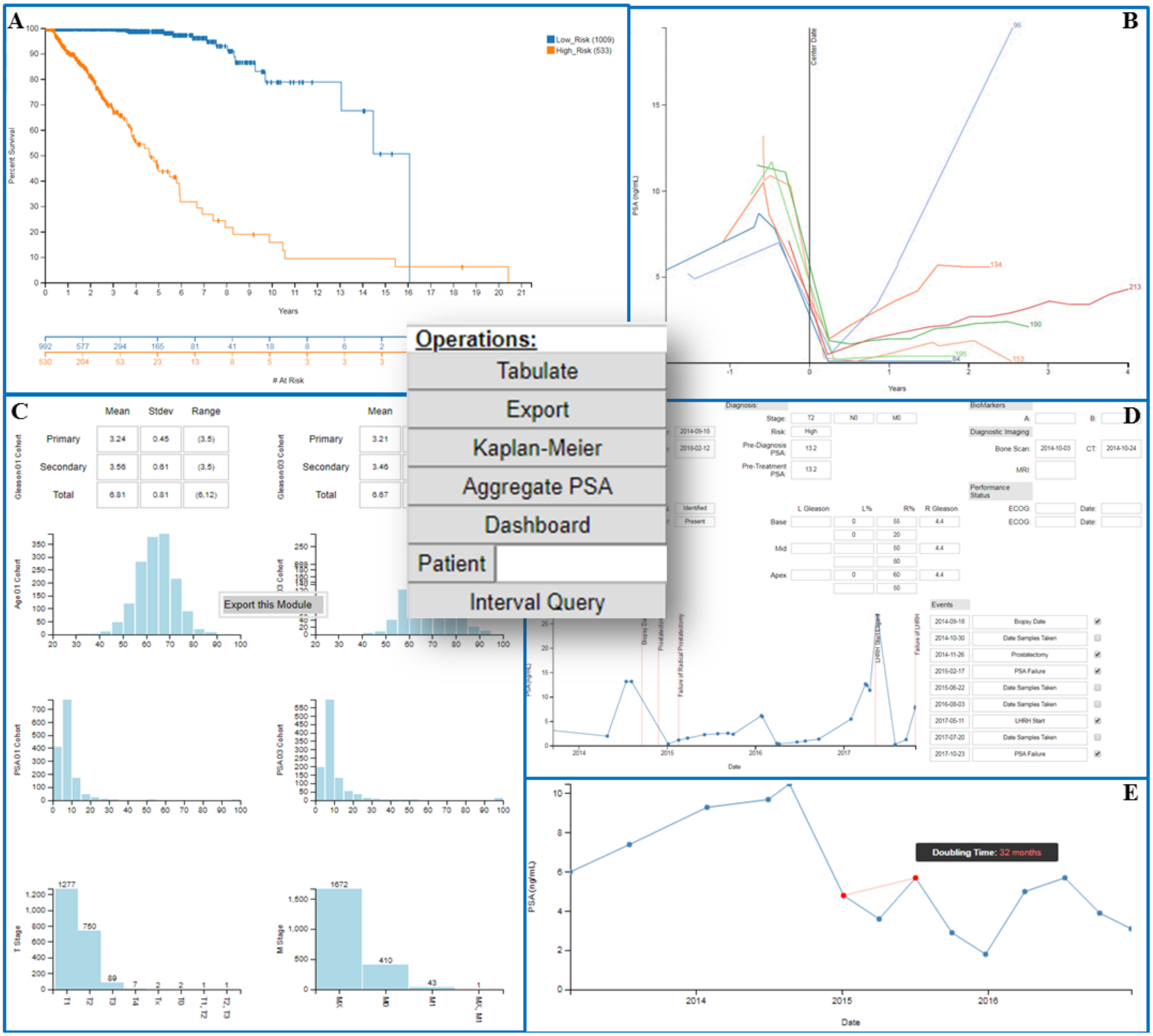
Results Pane

Rows per page: 20 Rows 1-20 of 540

Patient ID	Definition	Total gleason score (overall)	Clinical staging (T)	PSA
8	High_Risk	9	T1	7.8
9	High_Risk	8	T1	13
11	High_Risk	8	T2	171
13	High_Risk	8	T2	13.2
43	High_Risk	9	T2	11.1
46	High_Risk	9	T2, T3	2.4
70	High_Risk	9	T1	4.9
71	High_Risk	9	T1	6.2
76	High_Risk	7	T1	21.2
96	High_Risk	9	T2	7
103	High_Risk	9	T2	283
111	High_Risk	7	T2	23.3
112	High_Risk	9	T2	135.5
113	High_Risk	9	T2	14.3
119	High_Risk	7	T3	13

Label user-created definition as "High-Risk". Click on save and the definition is stored under Guest's Saved Definitions in the Fields Pane for future sessions.

Fig. 3.



		Mean	Stdev	Range
Gleason 01 Cohort	Primary	3.24	0.45	(3,5)
	Secondary	3.56	0.61	(3,5)
	Total	6.81	0.81	(6,12)

		Mean
Gleason 03 Cohort	Primary	3.21
	Secondary	3.46
	Total	6.67

2014-09-18
2018-02-12
Identified
Present

Diagnosis:

Stage: T2 N0 M0
Risk: High
Pre-Diagnosis PSA: 13.2
Pre-Treatment PSA: 13.2

	L Gleason	L%	R%	R Gleason
Base		0	55	4,4
		0	20	
Mid			50	4,4
			80	
Apex		0	60	4,4
			50	

BioMarkers

A: B:

Diagnostic Imaging

Bone Scan: 2014-10-03 CT: 2014-10-24
MRI:

Performance Status

ECOG: Date:
ECOG: Date:

Events

2014-09-18	Biopsy Date	☑
2014-10-30	Date Samples Taken	☐
2014-11-26	Prostatectomy	☑
2015-02-17	PSA Failure	☑
2015-06-22	Date Samples Taken	☐
2016-08-03	Date Samples Taken	☐
2017-05-11	LHRH Start	☑
2017-07-20	Date Samples Taken	☐
2017-10-23	PSA Failure	☑

Fig. 4.

Table 1. Comparison of PROSPeCT with related digital tools

Feature	Spreadsheet programs	Laboratory information management systems	Expert Systems (ES)	Electronic Medical/Health Record Systems (EMR/EHR)	Natural Language Processing (NLP) of EMR	Visual data-discovery Dashboards	Clinical Information Systems (CIS)	
Example	Microsoft Excel	Clarity LIMS[21]	MYCIN[21]	GE Centricity[8]	Algorithm[14]	InSightive Analytics[22]	REDCap[19]	PROSPeCT
Description	Desktop-based graphical user interface	Automate workflows, instruments, manage samples, data	Identify severely infectious bacteria, suggest effective therapies with a predefined model	Storage of patient data for records, billing, appointments, limited analysis	Automated PCa risk stratification data extraction method from EMRs by NLP	Custom-built, interactive dashboards; user examines data for patterns/trends	Server-based GUI for research databases	Secure web app designed to query and visualize data
Ease of Use								
Data storage	Any data type	Lab/literature data	Data for model	Any data type	Data stored in EMR/EHR	Any data type	Any data type	Any data type
Data handling-scale	Not designed for large data files	Efficient for few features/ large databases	Efficient for few features/ large databases	Efficient for few features/ large databases	Efficient for limited applications	Efficient for many features/ large databases	Efficient for few features/ large databases	Efficient for many features/ large databases
Database updated	Manual	Variable	Variable	Point of care	Automated	Automated	Automated	Automated
Interface	Manual	Variable	Questions prompt user to input data	Questions prompt user to input data	Questions prompt user to input data	Wizard guided	Designed for data capture	Drag & Drop, intuitive
Simple definition	Easy for user	Easy for user	Limited to built-in model	Variable/Easy for user	Limited to built-in application	Easy for user	Easy for user	Easy for user
Complex definition	Difficult, code writing needed	n/a	Limited to built-in model	n/a	Limited to built-in application	n/a	n/a	Easy, Boolean logic capacity
Time spent: query	High: ≥20 min	n/a	Med.: 10-20 min	Med.: 10-20 min	Med.: 10-20 min	Low: ≤ 10 min	n/a	Low: ≤ 10 min
Generation and Visualization of PCa patient results								
Single patient summary	Hard, write code	n/a	Variable	Easy, built-in app	Easy, built-in app	External analysis	n/a	Easy, built-in app
KME	Hard, write code	n/a	n/a	n/a	n/a	External analysis	n/a	Easy, built-in app
PSA trajectory plot	Hard, write code	n/a	n/a	n/a	n/a	External analysis	n/a	Easy, built-in app
PSA Doubling Time	Hard, write code	n/a	n/a	n/a	Easy, built-in app	External analysis	n/a	Easy, built-in app
Security								
Permissions	File-level password protection	Variable	Variable	Multi-tiered access levels to specific sections	Variable	Multi-tiered access levels to specific sections	Multi-tiered access levels to specific sections	Multi-tiered access levels to specific sections

n/a: not applicable to this program or platform.

References

1. Stewart BW, Wild CP, eds: World Cancer Report 2014. Geneva, Switzerland, World Health Organization, 2015

2. Canadian Cancer Statistics Advisory Committee. Canadian Cancer Statistics. A 2018 Special Report on Cancer Incidence by Stage. Toronto, ON, Canada, Canadian Cancer Society, 2018

3. Shariat SF, Karakiewicz PI, Roehrborn CG, et al: An updated catalog of prostate cancer predictive tools. Cancer 113:3075-3099, 2008

4. Evans SM, Millar JL, Wood JM, et al: The Prostate Cancer Registry: Monitoring patterns and quality of care for men diagnosed with prostate cancer. BJU Int 111: E158-166, 2013 (4 Pt B)

5. Burnett AL: Racial disparities in sexual dysfunction outcomes after prostate cancer treatment: Myth or reality? J Racial Ethn Health Disparities 3:154-159, 2016

6. Jayadevappa R, Chhatre S, Johnson JC, et al: Variation in quality of care among older men with localized prostate cancer. Cancer 117:2520-2529, 2011

7. Schroeck FR, Kaufman SR, Jacobs BL, et al: Regional variation in quality of prostate cancer care. J Urol 191:957-962, 2014

8. Knighton AJ, Belnap T, Brunisholz K, et al: Using electronic health record data to identify prostate cancer patients that may qualify for active surveillance. EGEMS (Wash DC) 4:1220, 2016

9. Courtwright AM, Gabriel PE: Clinical databases for chest physicians. Chest 153:1016-1022, 2018

10. Horvath MM, Rusincovitch SA, Brinson S, et al: Modular design, application architecture, and usage of a self-service model for enterprise data delivery: The Duke Enterprise Data Unified Content Explorer (DEDUCE). J Biomed Inform 52:231-242, 2014

11. Horvath MM, Winfield S, Evans S, et al: The DEDUCE Guided Query tool: Providing simplified access to clinical data for research and quality improvement. J Biomed Inform 44:266-276, 2011

12. Nigrin DJ, Kohane IS: Data mining by clinicians. Proceedings/AMIA: Annual Symposium:957-961, 1998

13. Huser V, Narus SP, Rocha RA: Evaluation of a flowchart-based EHR query system: A case study of RetroGuide. J Biomed Inform 43:41-50, 2010

14. Kim C-S, Lee JY, Chung BH, et al: Report of the Second Asian Prostate Cancer (A-CaP) Study Meeting. Prostate Int 5:95-103, 2017

15. e-PROSPeCT. https://github.com/max-uhlich/e-PROSPeCT.

16. Alberta Prostate Cancer Research Initiative (APCaRI): Alberta Prostate Cancer Research Initiative homepage. https://apcari.ca.

17. PostgreSQL Global Development Group: PostgreSQL: the world's most advanced open source relational database. https://www.postgresql.org/.

18. van Rossum G: Python tutorial. Technical Report CS-R9526. Amsterdam, the Netherlands, Centrum voor Wiskunde en Informatica. Amsterdam, 1995

19. Goel MK, Khanna P, Kishore J: Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res 1:274-278, 2010

20. National Cancer Institute: SEER data & software. https://seer.cancer.gov/data-software.

21. American College of Surgeons, American Cancer Society: National Cancer Database. https://www.facs.org/quality-programs/cancer/ncdb.

22. University of California, San Francisco, Department of Urology: CaPSURE. https://urology.ucsf.edu/research/cancer/capsure.

23. Gregg JR, Lang M, Wang LL, et al: Automating the determination of prostate cancer risk strata from electronic medical records. JCO Clin Cancer Inform [epub ahead of print on June 8, 2017]

24. Febbo PG, Mulligan MG, Slonina DA, et al: Literature Lab: A method of automated literature interrogation to infer biology from microarray analysis. BMC Genomics 8:461, 2007

25. Varian Medical Systems: InSightive analytics. https://www.varian.com/oncology/products/software/information-systems/insightive-analytics.

26. Harris PA, Taylor R, Thielke R, et al: Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 42:377-381, 2009

Supplementary Material

The use of the PROSPeCT interface was approved by The Health Research Ethics Board of Alberta Cancer Committee, HREBA.CC, through study numbers HREBA.CC-26198 and HREBA.CC-14-0085.

Alberta Prostate Cancer Registry

Prostate cancer patients are recruited to join the Alberta Prostate Cancer Registry hosted by the Alberta Prostate Cancer Research Initiative (APCaRI) on an ongoing basis. Clinical research teams then collect patient data at the time of consent and thereafter, as the patient undergoes testing, biopsy, and possible treatment. Patient data includes demographic information, co-morbidities, medications, family history of PCa, PSA values, PCa diagnosis Gleason scores, treatments received, disease-specific outcomes, and quality of life self-reported outcomes. The standard-of-care clinical data from each enrolled patient is updated on average every six months into the Alberta Prostate Cancer Registry, which is managed by the REDCap (Research Electronic Data Capture) web application hosted at the University of Alberta[19]. REDCap is a secure web application designed to support data capture for research studies, providing: 1) an interface for validated data entry; 2) audit trails for tracking data manipulation and export procedures; 3) export procedures for data downloads to common statistical packages; and 4) procedures for importing data from external sources.

Python modules: Gen_Define, Gen_Populate, Gen_Derive, Gen_XML, and Remote_Ops

The Gen_Define module combs the datacut from REDCap and builds PostgreSQL table statements for use in the database underlying PROSPeCT. The fields defined by this module are guaranteed to conform to some data regularity conditions required for the next modules, such as uniformity of datatype and non-emptiness. This module also handles various special cases which exist in REDCap but require conversion for PROSPeCT. For example, some fields in REDCap are categorical, meaning the field itself contains integer values, but each integer maps into a set of categories which are expressed as a sequence of characters (strings). For instance, a given row might contain an integer such as 1 or 2 or 3, where 1='Benign', 2='Adenocarcinoma', and 3='Other'. In special cases, a field with three possibilities is expanded into three Boolean fields, one for each category. Each of these Boolean fields has a value of 1 to indicate that the record belongs to a respective category and a value of 0 to indicate that the record does not belong to that category. Among many other operations, The Gen_Define module converts these types of fields into a single string field for use in PROSPeCT. This component takes approximately 26 seconds to complete.

The Gen_Populate module uses the previous modules database definition as its guide while creating PostgreSQL insertion statements for every row in the database. Every patient in the database can have one or more rows in the datacut, where each column is a REDCap field. This module parses every row of the datacut and generates at most one insertion statement per table. This operation is completed in around 42 seconds.

PROSPeCT currently provides ten fields computed from various REDCap fields. The Gen_Derive module facilitates the creation and population of the derived tables and their fields. There is one Python function for every derived field, accepting the relevant parameters and returning a specific result. These parameters are extracted from REDCap and then sorted, consolidated, grouped, and sent to their respective functions according to the mappings expressed in a set of comma-separated value (csv) files. Gen_Derive also passes a modified database definition, including the new derived fields and tables, to the next module to be referenced as the final state of the PROSPeCT database. This component currently completes in around 18 seconds.

The Gen_XML module creates an .xml description of the final state of the dataset to be used by the PROSPeCT web application to define a programmatic connection to the underlying PostgreSQL database. The last run of this module completed in 2 milliseconds.

The Remote_Ops module facilitates all operations which must occur remotely on the PROSPeCT server. It has two subprocesses: ‘rebuild’ and ‘restart’. The first process connects to the remote server through a Secure Shell (ssh) protocol, destroys the current version of the dataset contained in PostgreSQL, and recreates it using the newly-generated statements created by the previous four modules. Any anomalous behaviour is recorded and passed back to the home computer. The second process connects to the remote server once again through ssh, but this time using a special administrative account. This process restarts the server and logs out.

As a patient privacy measure, the REDCap system automatically de-identifies the data, shifts all dates, then assigns each patient a unique APCaRI identification number before uploading the data to PROSPeCT. Any database entry errors identified by the Python modules or PROSPeCT users are reported to the APCaRI team for correction.

Results Pane

Each column can be:

- Removed
- Moved left
- Moved right

- Filtered: At a given time, only one column can be filtered, which will remove rows from the table. For example, a column with numerical values can be filtered based on the criterion "v >= 8", where "v" denotes the values present in that column (e.g., the Total Gleason score). This filter will remove any rows from the table for which that criterion is not satisfied.
- Sorted: A column can be sorted in ascending or descending order by clicking the "ascending" or "descending" buttons in the drop-down menu that appears after clicking on the header of a column.

Definition of terms

1. PCa risk stratification (Fig. S1 and Table S1): Prostate cancer risk stratification classification schemes are the most common way that physicians determine the aggressiveness of prostate cancer. Each risk group predicts how quickly the prostate cancer is likely to grow and/or to spread outside of the prostate. Recommendations regarding which forms or combination of forms of treatment are highly influenced by the risk category of a patient. This risk status is calculated by assessing results from prostate biopsy, PSA value, and clinical TNM staging through a digital rectal exam:

- Low Risk: T1- T2a and Gleason score ≤ 6 and pre-biopsy PSA <10 ng/mL.
- Intermediate: T2b-T2c or Gleason is 7 or pre-biopsy PSA 10-20 ng/mL.
- High Risk: T3a or higher or Gleason score ≥ 8 or pre-biopsy PSA >20 ng/mL.

2. PCa recurrence/progression (also known as biochemical recurrence or PSA failure): This denotes the state of an increasing prostate-specific antigen (PSA) level among men treated with

prostatectomy or radiation therapy for localized prostate cancer. This typically occurs when the disease is recurring (e.g., post-curative treatment) or progressing (e.g., growth of the cancer in the body). Disease Recurrence after Prostatectomy is a common definition employed in user queries, indicating that prostate cancer has progressed despite the treatment, and PROSPeCT can compute this. This definition requires a set of computations carried out offline by the import modules using multiple data fields from REDCap. As such, this definition is easy to use in PROSPeCT, but difficult to implement using tools like Excel. For example, the indicator of the simplest recurrence scenario (failure of prostatectomy) is defined as a pair of two consecutive PSA measurements (corresponding to index t and t+1): $PSA_t >= 0.2$ and $PSA_{t+1} >= 0.2$ occurring after prostatectomy. PSA Failure is computed for each of the following treatments:

- Post prostatectomy: After prostatectomy, any consecutive pair of values x>=0.2 followed by y>=0.2
- Post chemotherapy (curative): After having reached nadir (the last PSA before increases are documented), any value x>=(nadir*1.25)
- Post cryotherapy (primary): After having reached nadir (the last PSA before increases are documented), any value x>=(nadir+2)
- Post Androgen Deprivation Therapy: ADT (Enzalutamide, Flutamide, Abiraterone, Bicalutamide, or Other) - After having reached nadir (the last PSA before increases are documented), any value x>=(nadir*1.25) occurring between the start and stop dates of treatment. If a treatment does not have a stop date, we take the last PSA date, so as to include treatments which may be currently in progress.
- Post Luteinizing Hormone-Releasing Hormone: LHRH - (Eligard, Zoladex, Suprefact, or Degarelix): After having reached nadir, any pair of values x and y occurring between the start

and stop dates of treatment such that (x>nadir), (y>x), and (y-x >= 1 week). If a treatment does not have a stop date, we take the last PSA date, so as to include treatments which may be currently in progress.

- Post radiation therapy: (Curative) (HDR/LDR Brachy, External Beam): define bounce as only occurring while no other treatments are taking place. RT can fail in the presence of other treatments; however, it cannot bounce in their presence.
- A PSA rise was defined as an increase of at least 2.0 ng/mL in PSA level over a pre-rise nadir (nadir+2). If the PSA rise was followed by a decrease to <=0.5ng/mL without intervention (Hormones, ADT, Chemo, Cryo, Prostatectomy, non-curative Radiation), it was considered a benign bounce. Only one bounce allowed per patient. Failure conditions were as follows:

* if the PSA level >= nadir+2 and we hit an intervening treatment, this is a failure
* if the PSA level >= nadir+2 and we do not hit an intervening treatment, but we hit the end of our measurements, this is a failure
* if the PSA level >= nadir+2 but we fall to <=0.5 without hitting an intervening treatment or the end of our measurements, that was a bounce and we have a new nadir
* if the PSA level < nadir+2 and we hit a new treatment, keep going until the end of measurement
* if the PSA level < nadir+2 and we do not hit a new treatment, but the end of our PSA measurements is hit, this is not a failure.

Kaplan-Meier Estimator Parameters

Kaplan Meier curves are generated for each selected population. By default, the Start date is the “Date of Biopsy”, the end date is the date of “PSA Failure”, and the censor date is “Last

Contact Date”. The three parameters can be changed by dragging a date field into the relevant placeholder in a Kaplan Meier tab. PROSPeCT will display that Kaplan-Meier curve in the Results Pane once the user presses the *Recompute* button. The tab structure of the Results Pane can accommodate multiple Kaplan Meier curves.

- Start date: For example, this could be the “biopsy date”, which is the date of the patient first biopsy
- End date (or Failure date), which could be the date of the PSA Failure, if that happened for the patients
- Censor date: last contact date represents the date when the patient was last contacted (e.g., the date when the participant left the study)

Survival Rate Graph:

- Log rank statistics: p-value for comparing the survival probabilities of the populations displayed in the KM plot
- X-axis: years
- Y-axis: probability of patient survival for the time on the x-axis
- Under the graph, there is an “At Risk” timeline for each definition that includes the number of patients who are “at risk” with regards to the event, corresponding to a time point on the x-axis.

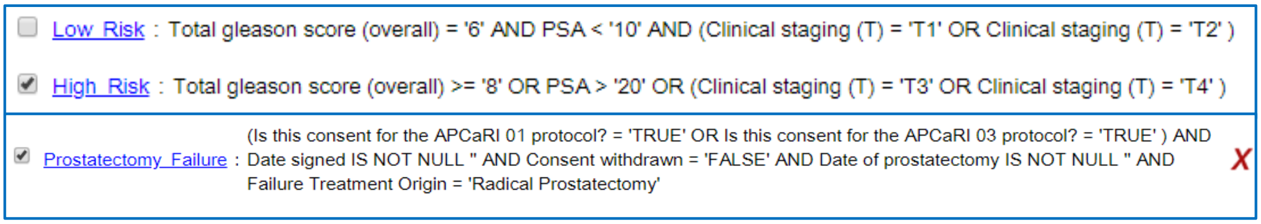
Aggregate PSA Plot Parameters

PSA aggregate graph provides information about the patient’s PSA values over time, as shown in Fig. 4. The x-axis represents the years relative to the event (i.e., x = 0; the centering date) when the biopsy was performed, while the y-axis represents the PSA values measured in nanograms per milliliter (ng/mL).

- Center date. This could be any date field by drag-&-drop into the date placeholder; the default is the Biopsy date (e.g., the patient's first biopsy). PROSPeCT will generate a new graph when the user presses the *Recompute* button. This graph displays the aggregate PSA trajectories for a defined population. The PSA trajectories are centered around a specific date and displayed on the same axes. The default center date is 'Date of Biopsy'. This allows a general view of how a treatment or other event can affect the PSA trajectories of a large population.

Table S1. A sampling of fields available in PROSPeCT (in specific categories).

Primary data fields directly imported from REDCap	
Patient information (up to # fields)	
Birthdate	Either actual date or time-shifted depending on user's access level.
Body measurements	Height and weight.
Medications and Supplements (up to # fields)	
PSA (up to # fields)	
PSA levels	Time-stamped PSA level measurements, pre and post-biopsy.
Imaging (up to # fields)	
Pathology (up to # fields)	
Diagnostic biopsy	Biopsy dates, prior procedures, digital rectal exam (DRE) findings, transrectal ultrasound (TRUS) findings, gland volume, PSA density value, Grade group, total cores collected, number and proportion of tumour-involved core tissues, perineural invasion and location, periprostatic fat invasion and location, high grade PIN, cancer type, Cribriform pattern, intraductal carcinoma, needle core biopsies.
Clinical staging:	T1, T1a-c, T2, T2a-c, T3, T3a, b to describe the size and location of tumour.
Gleason grades (primary + secondary) and Grade group	Gleason grades 3 + 3 = 6; Grade group 1 Gleason grades 3 + 4 = 7; Grade group 2 Gleason grades 4 + 3 = 7; Grade group 3 Gleason grades 4 + 4 = 8; Grade group 4 Gleason grades 4 + 5, 5 + 4, 5 + 5 = 9-10; Grade group 5
TURP	Transurethral resection of the prostate
Radical prostatectomy	Time-stamped with date
Post-treatment biopsy	Time-stamped with date
Treatment (up to # fields)	
Radiation therapy	Time-stamped with date
Progression (up to 27 fields)	
Quality of life questionnaires (up to # fields)	
REDCap-derived fields: computed offline during data transfer from REDCap to PROSPeCT	
Prostate cancer risk stratification	Low risk = Clinical stage T1-T2a AND Gleason score <= 6 AND pre-biopsy PSA < 10 Intermediate Risk = Clinical stage T2b-T2c OR Gleason score = 7 OR pre-biopsy 10 <= PSA <= 20 High risk = Clinical stage T3a or higher OR Gleason score >= 8 OR pre-biopsy PSA > 20
Prostate cancer recurrence/progression (also termed biochemical recurrence or PSA failure)	PROSPeCT currently provides PSA Failure values for patients who have undergone radical prostatectomy, chemotherapy, radiation therapy, cryotherapy, antiandrogen therapy, and LHRH therapy.
On-the-fly-derived fields computed by PROSPeCT	
Patient age	Age at time-stamped events computed from the patients' birthdate.
PSA doubling time	The time it took for the patients' PSA levels to double. Calculated from the PSA trajectory plot in a single patient summary.
Date Intervals	Date intervals can be computed on-the-fly between any two date fields using the Interval Query operation.



□ Low_Risk : Total gleason score (overall) = '6' AND PSA < '10' AND (Clinical staging (T) = 'T1' OR Clinical staging (T) = 'T2')

☑ High_Risk : Total gleason score (overall) >= '8' OR PSA > '20' OR (Clinical staging (T) = 'T3' OR Clinical staging (T) = 'T4')

☑ Prostatectomy_Failure : (Is this consent for the APCaRI 01 protocol? = 'TRUE' OR Is this consent for the APCaRI 03 protocol? = 'TRUE') AND Date signed IS NOT NULL '' AND Consent withdrawn = 'FALSE' AND Date of prostatectomy IS NOT NULL '' AND Failure Treatment Origin = 'Radical Prostatectomy'

Fig. S1. The Definition pane enables users to create their own custom definitions, for example "Low Risk", "High Risk", and "Prostatectomy Failure" to identify a patient cohort with disease progression.