Pilot 6: Prostate Cancer

1. Key Information

1.1 Involved Partners

- Philips Electronics Nederland B.V. (PHI)
- Stockholms Lans Landsting (KAR)

1.2 Involved Countries

- Netherlands
- Sweden

1.3 Keywords

- Clinical decision support
- Medical risk prediction
- Treatment risk prediction
- Support of the surgery strategy decision making process

1.4 Task Description

Surgery is one of the main treatment options for prostate cancer today. There are multiple aspects that need to be considered when planning for the removal of the prostate. On the one hand, the oncological control of the tumor is the most relevant to ensure as much as possible that all cancer has been removed and does not return during follow-up. On the other hand, the aggressive removal of all prostate structures, e.g. nerve bundles will likely lead to poor functional outcomes like urinary incontinence of sexual dysfunctions. Consequently, the appropriate balance between oncological and functional surgery outcome is of utmost relevance to the patient. This pilot aims to support the decision-making process of how the surgery should be performed in order to provide the most optimal balance between tumor control and urological function after treatment. To this end, multiple risk models are provided based on the integration of heterogeneous data like demographics, laboratory, urology, oncology, pathology, radiology, nursing, health economics as well as data from patients themselves.

Led by Karolinska University Hospital (KAR), pilot 6 has designed a Clinical Decision Support (CDS) system, called ISPM Prostate, which has become an officially released product for clinical use. ISPM is installed on premise, on a windows server VM administered and controlled by the hospital IT department. ISPM consists of angular/typescript frontend and a java micro services backend, running on a local web server. Data is stored in a local FHIR database. To install/update software, the VM can be accessed using a secure VPN connection set up by hospital IT. ISPM runs completely locally inside the hospital firewall. OncoPredict is a separate research prototype developed in BigMediyltics, also installed on the VM. It retrieves de-identified data from ISPM, to execute statistical analysis. The results of the analysis are then visually processed and presented to the clinical staff.

Pilot 6 also aims to enhance patient outcomes and increase productivity in prostate cancer care with curative intent, by developing random forest classifiers using the complex datasets while ensuring security and privacy of personal data. The CDS and the captured big data will be used to derive primary treatment decisions in a multidisciplinary setting, derive treatment and value-based health care-related quality outcome measures, create decision models to improve functional outcome predictions after primary intervention and apply health economic modelling to test cost-effectiveness of the implemented big data technologies.

2. Building Blocks

2.1 Architecture

2.1.1 System Architecture

Software design
Led by Karolinska University Hospital (KAR), this pilot has designed a Clinical Decision Support (CDS) system. The CDS platform consists of the configurable IntelliSpace Precision Medicine (ISPM) software from Philips. The CDS is used to deliver primary treatment decisions in a multidisciplinary setting, derive treatment and value-based health care-related quality outcome measures, create decision models to improve functional outcome predictions after primary intervention and apply health economic modelling to test cost-effectiveness of the implemented big data technologies. The ISPM Prostate environment has been integrated to hospital information systems via the dedicated research data warehouse at KAR (KarDa).

The diagram below shows the ISPM Prostate application, which consists of an HTML/JavaScript web client that interacts with several web services via REST interfaces. The component labelled as 'R' indicates the language and environment for statistical computing R in which risk models are exposed, sending patient information and returning risk predictions using the REST API.

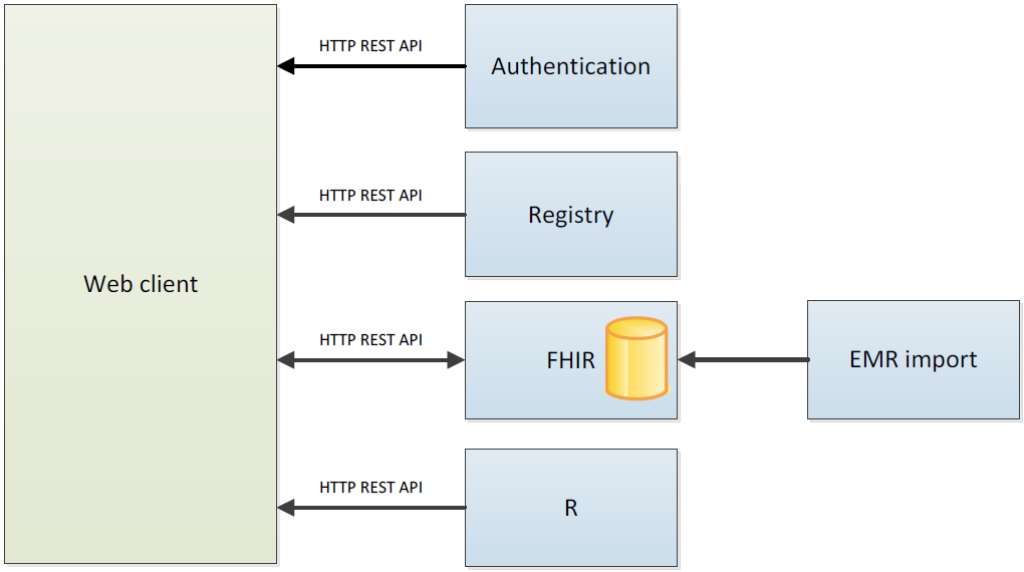

Image 1. ISPM architecture

Data entry and integration in Karolinska University Hospital
The diagram below shows the various data import and export pipelines that extract, transform and load data into the specific databases. The systems Orbit, Sympathy and TakeCare are hospital information systems at KAR from where comprehensive databases are collected into the data warehouse KarDa. KarDa hosts a Microsoft SQL server such that ISPM can retrieve relevant information using SQL queries. Currently under development is the ability to send or

export the structured data in ISPM to the swedish cancer registry NPCR. INCAnet is developing a system that exposes a RESTful API that allows to send patient data directly to the NPCR automatically.

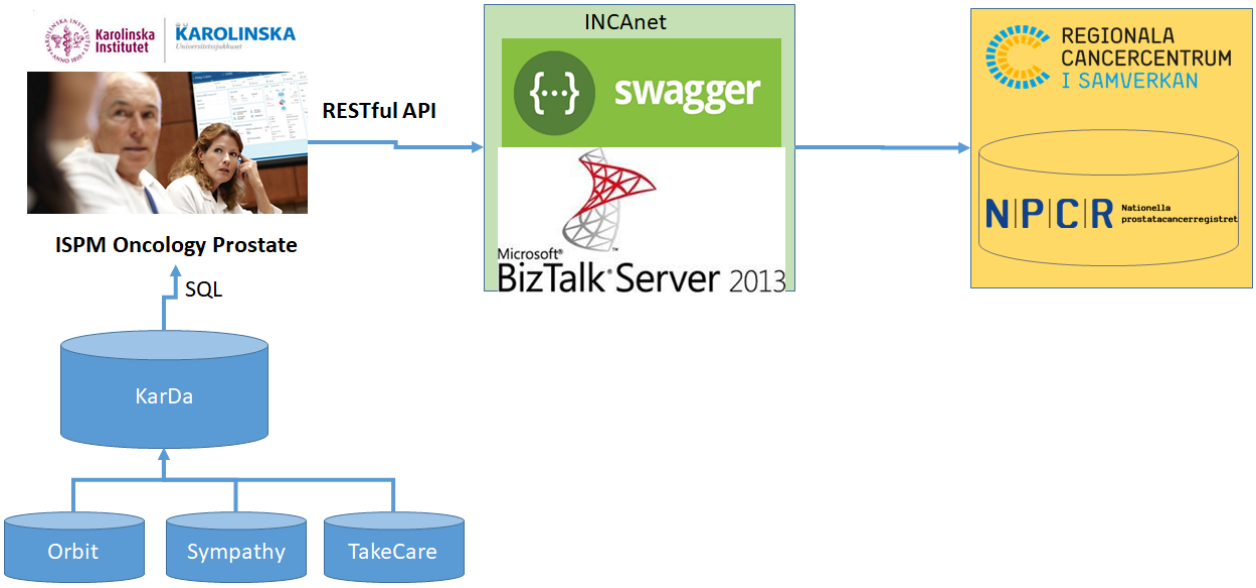

Image 2. Data flow in pilot 6

Implemented Features of the ISPM Solution

Application views

The miProstate KAR application consists of three different views (visualized below):

1. Patient worklist
2. Manual input forms
3. MDT dashboard

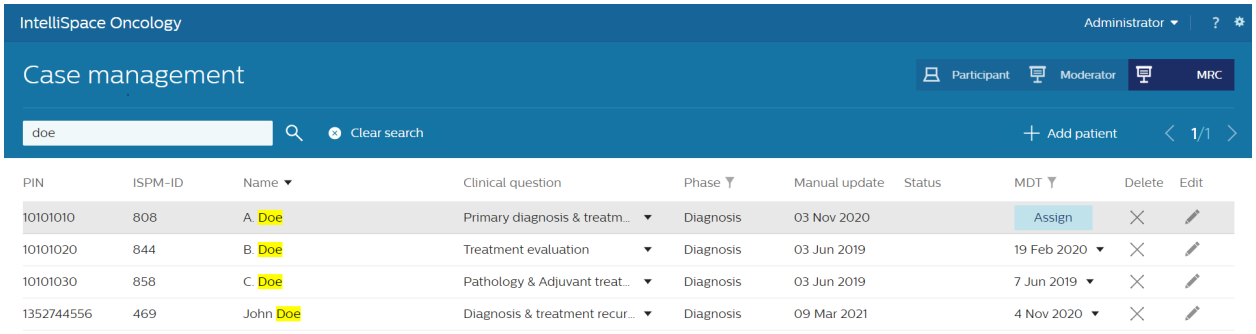

Image 3. Patient worklist with selected parameters visualized and interactive features for patient management, such as selection of view type, importing (Add patient), sorting, selecting, searching, assignment of clinical question, assignment to MDT etc. and a hyperlink to the edit sections for each patient.

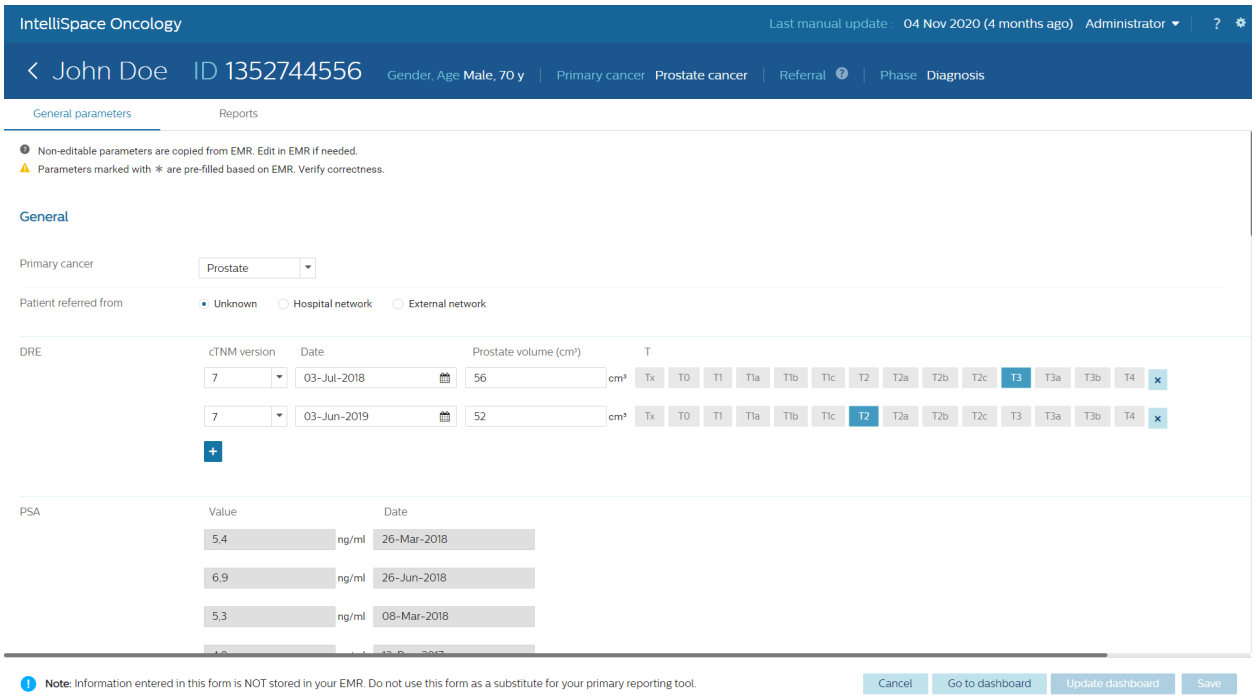

Image 4. Input section of general parameters

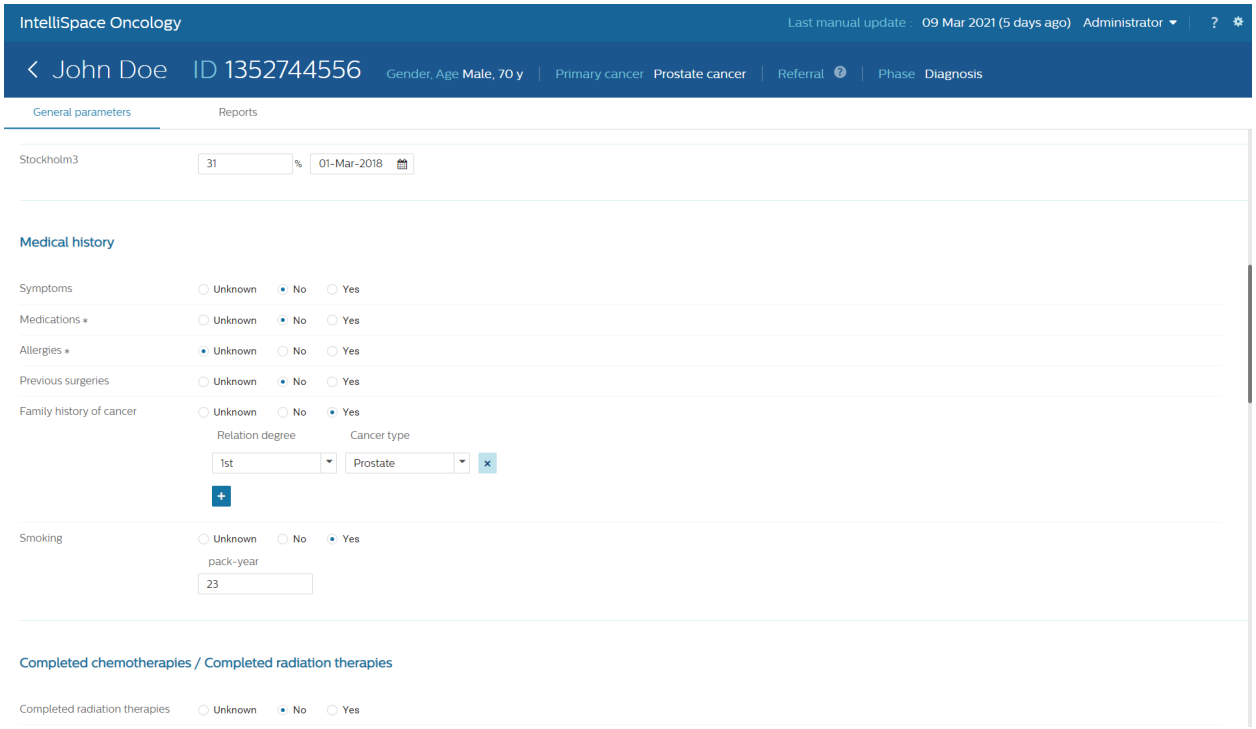

Image 5. Cont. Input section of general parameters

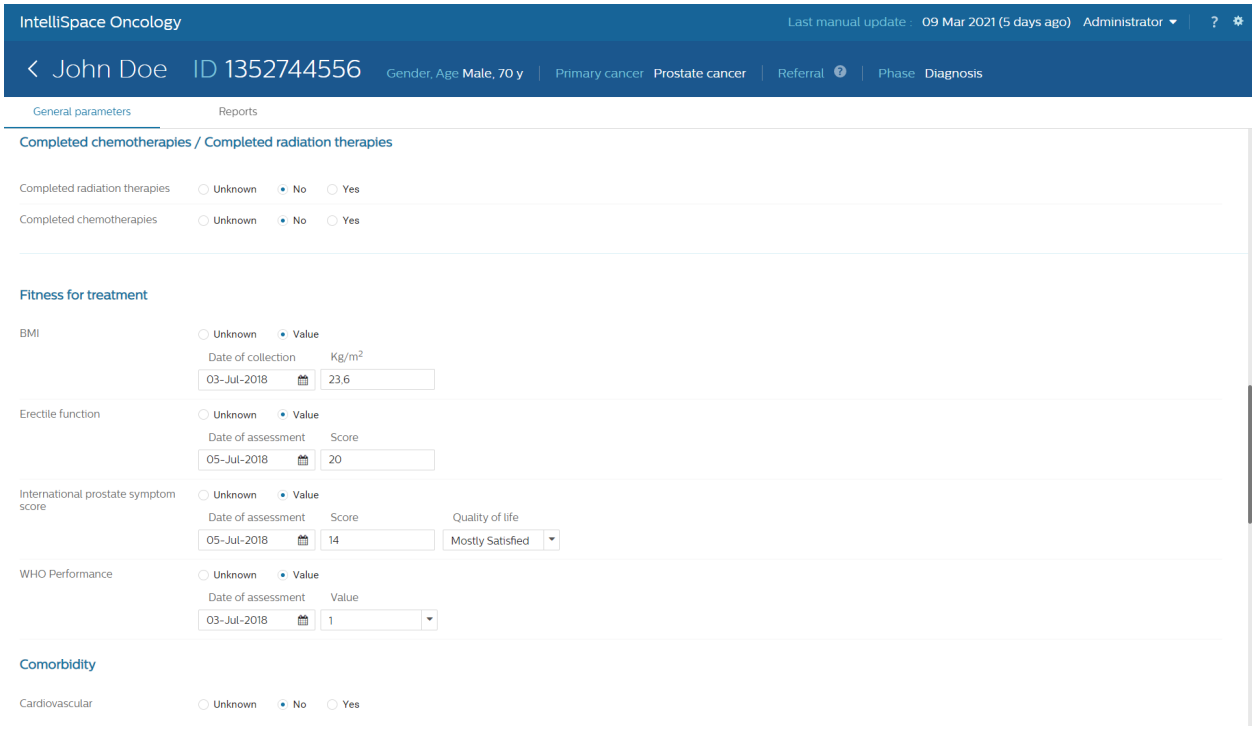

Image 6. Cont. Input section of general parameters

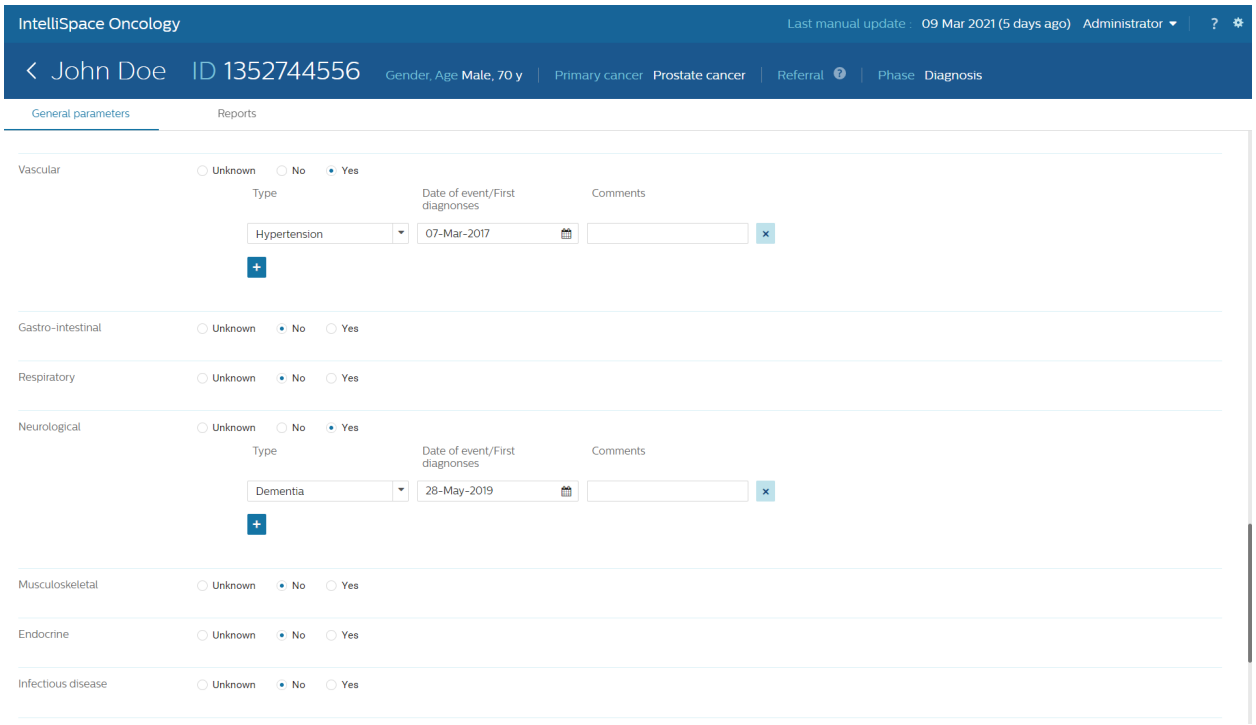

Image 7. Cont. Input section of general parameters

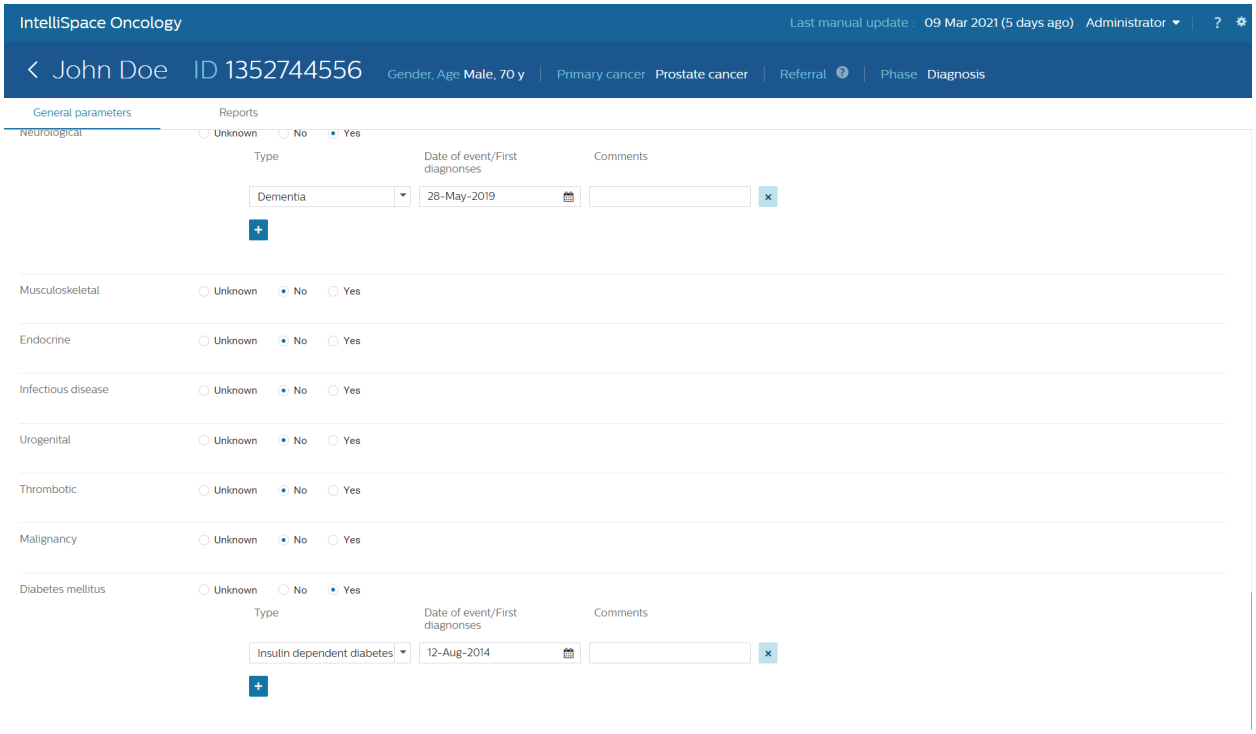

Image 8. Cont. Input section of general parameters

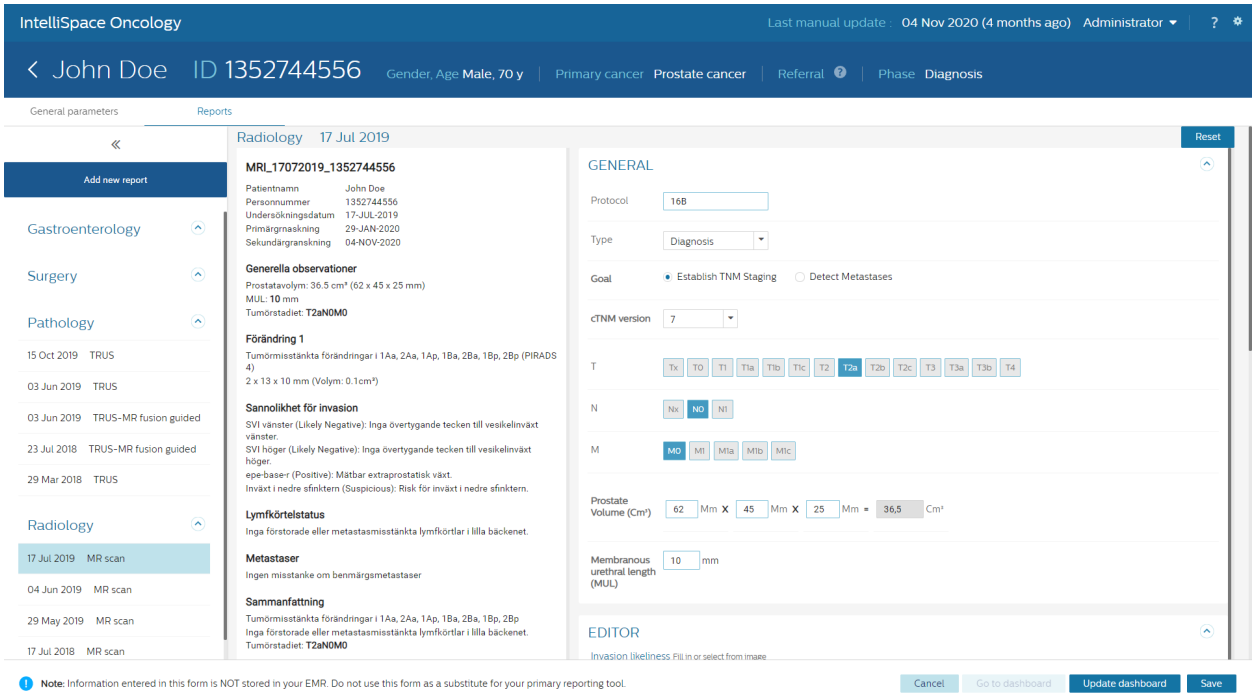

Image 9. Input section of radiology parameters. Radiology report generated by ISPM in the second column from the left.

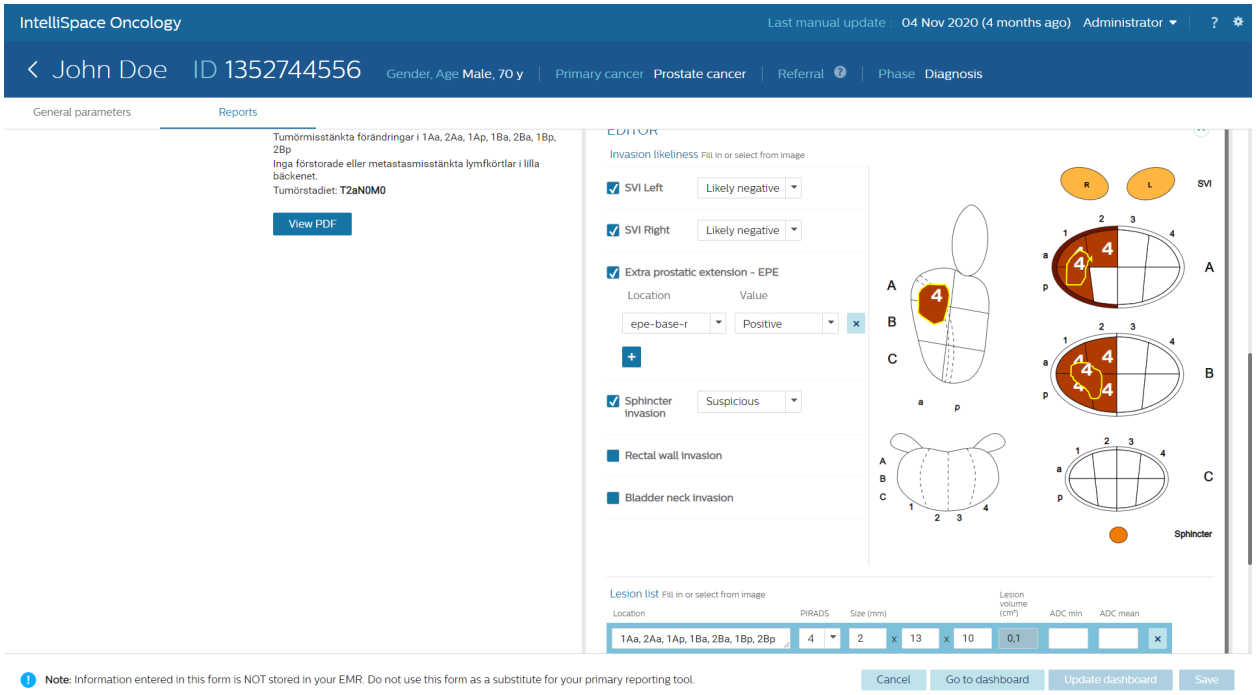

Image 10. Input section of radiology parameters for visualization of lesion location

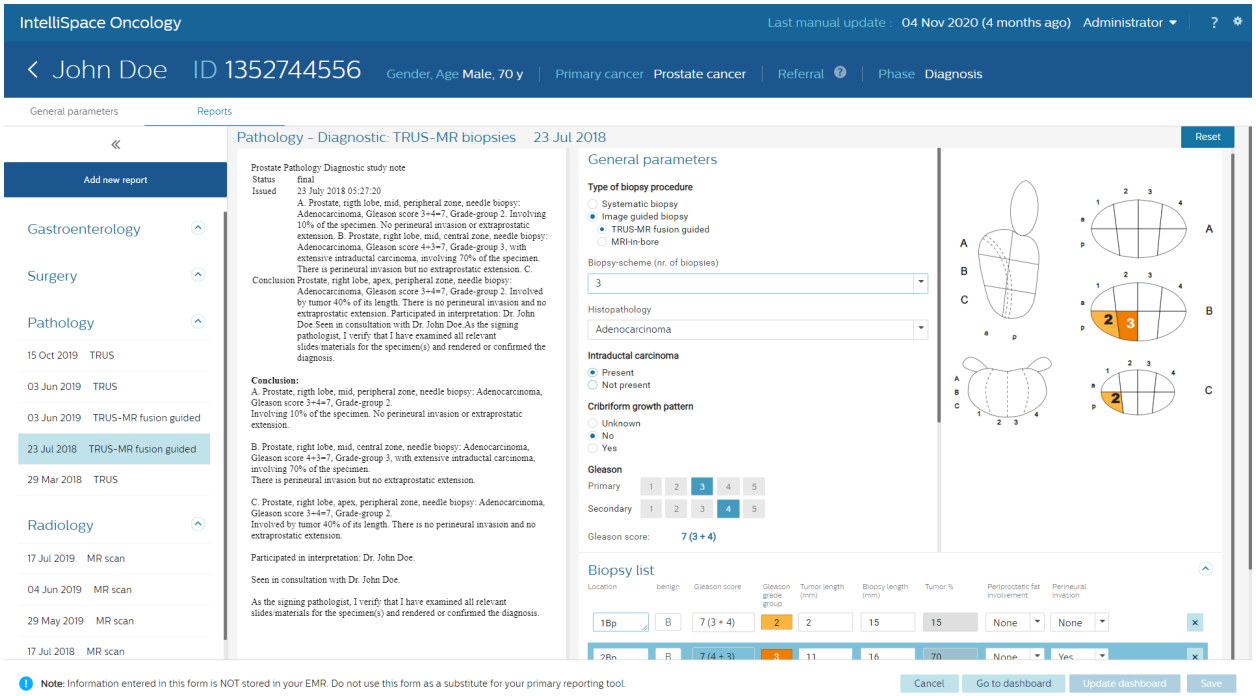

Image 11. Input section of biopsy parameters. Biopsy report automatically retrieved from KarDa in the second column from the left.

MDT dashboard

The MDT dashboard shows the relevant clinical data for the selected patient. This view is used in the MDT setting, where treatment strategy is decided.

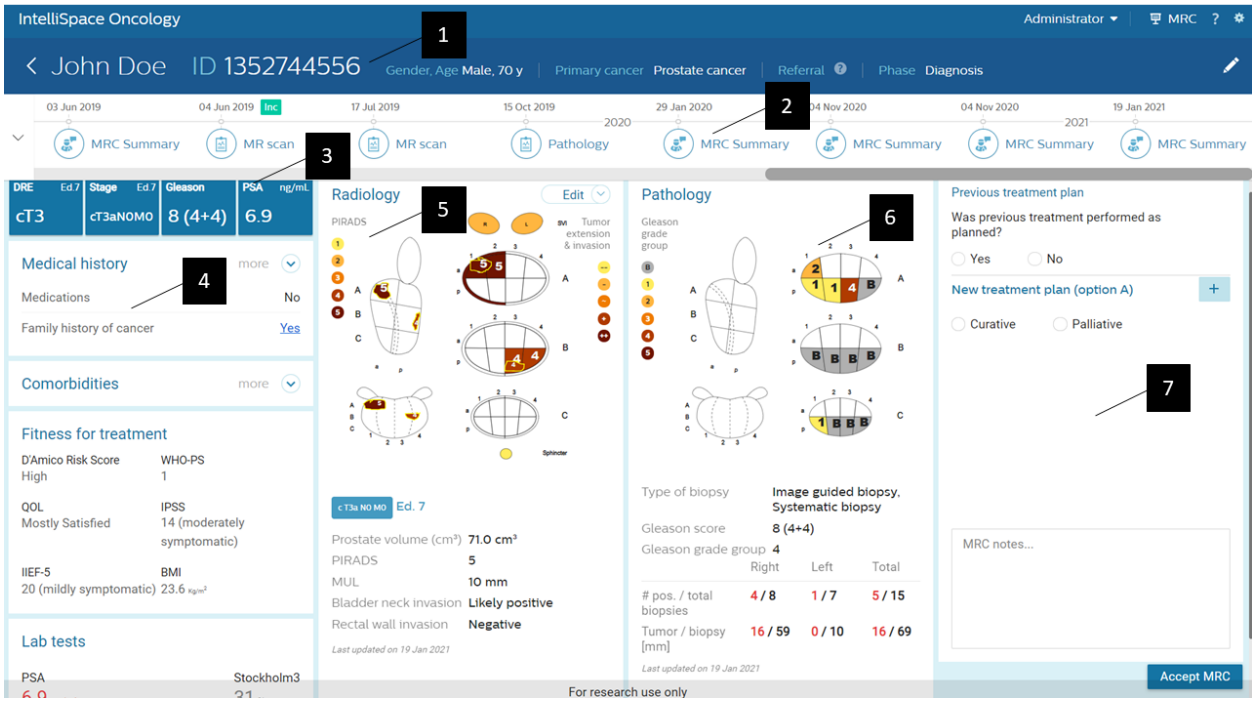

Image 12. MDT dashboard view of ISPM

1. Patient bar: shown at the top of the page. It identifies the selected patient and shows summary info.
2. Diagnosis bar: provides an overview of the relevant diagnostic events of the selected patient.
3. Diagnosis summary: provides an overview of core diagnostic parameters
4. General tab: provides an overview of general parameters, history and patient-reported data
5. Radiology tab: presents a visualization of radiological information and lists radiological data.
6. Pathology tab: presents a visualization of pathology information and lists pathology data
7. Treatment selection tab: is used to record the treatment selection.

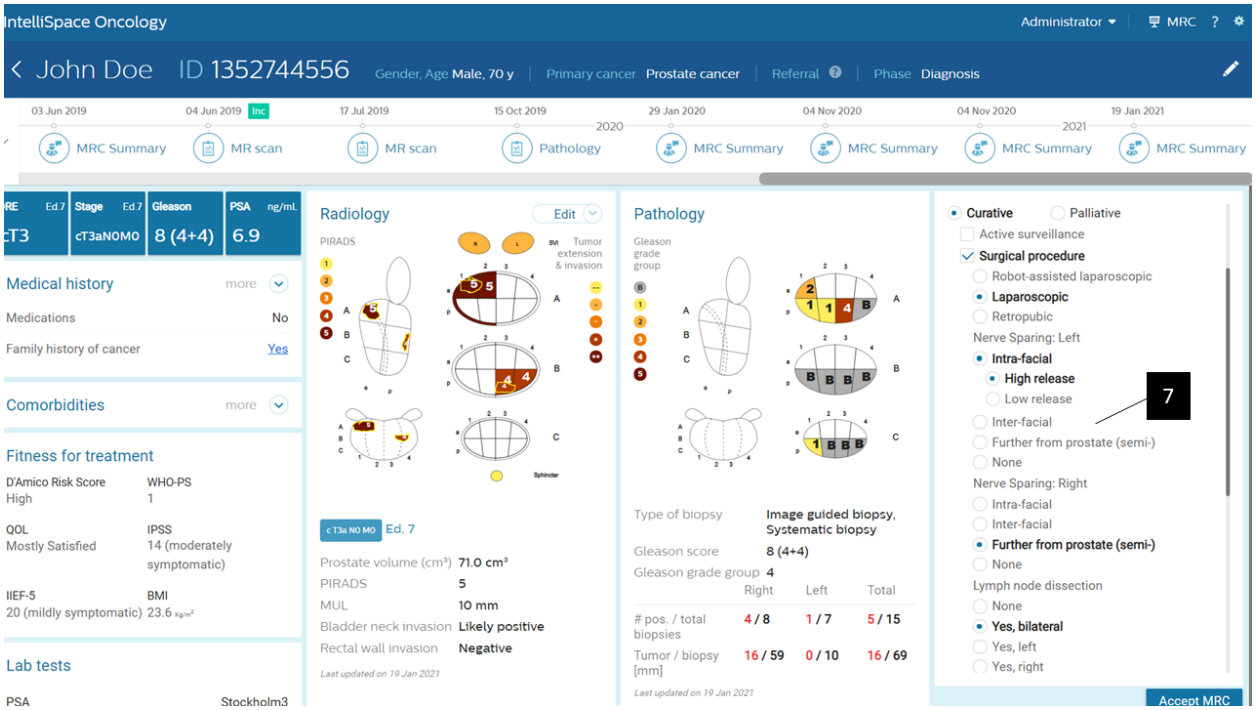

Image 13. Cont. MDT dashboard view. 7. Treatment selection tab: Example: Surgery strategy is selected in an expandable tree.

2.1.2 Data Flow & Interoperability of services

In the present study setting, a selection of patient data is automatically imported from KarDa upon entering of the patient PIN in the “Add patient” field in the patient list in ISPM and subsequently stored in the FHIR database, which is further enriched by manual data entry in the input sections of ISPM.

2.1.3 Necessary Hardware

No additional hardware is required. ISPM and OncoPredict are installed on a VM in existing KAR servers.

2.1.4 Software Components

ISPM and OncoPredict are installed on a VM in existing KAR servers.

2.3 Data Processing

2.3.1 Processing of large structured / unstructured data sources

2.3.1.1 Data Sources

To support the assessment of the productivity gains we aim to implement the ISPM Prostate environment with an integration to hospital information systems via the dedicated research data warehouse at KAR (KarDa). Direct integration to live clinical IT systems at KAR is restricted to clinically approved software only. An envisioned solution is shown below (Image 14).

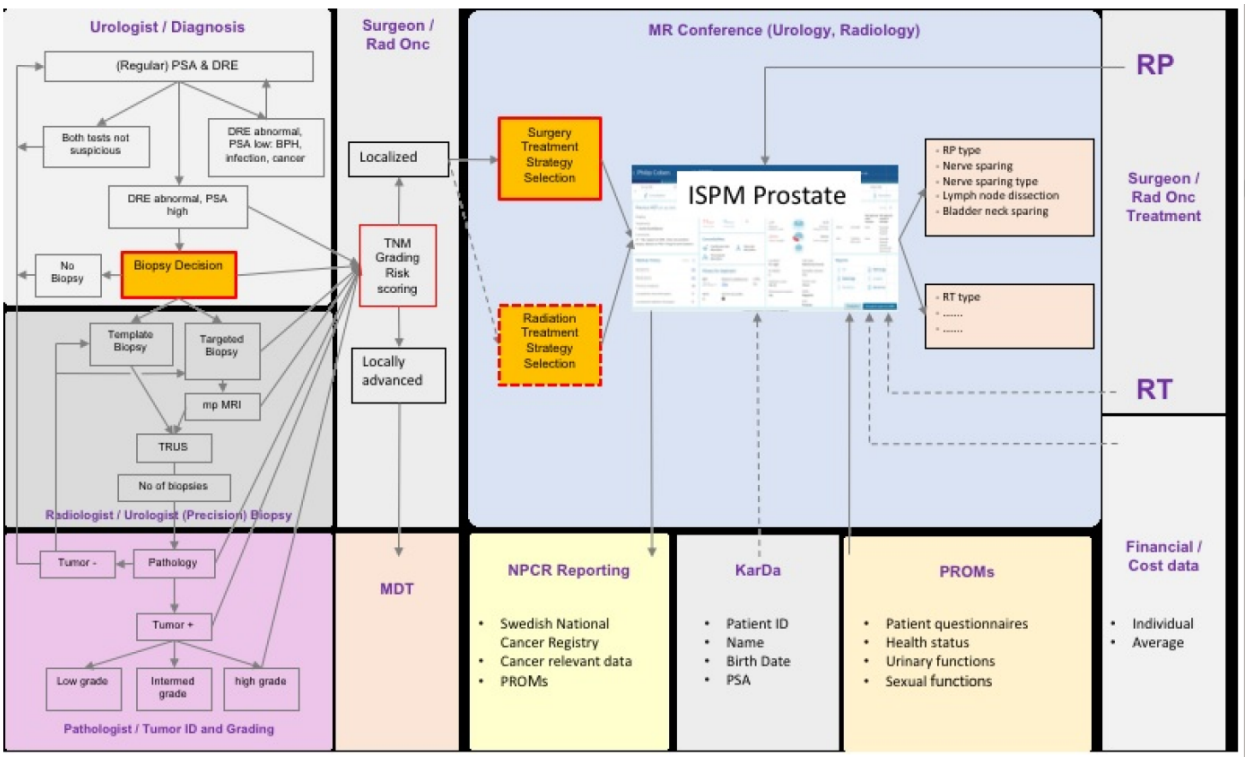

Image 14. Pilot 6 Flow chart of ISPM Prostate CDS for primary, localized Prostate Cancer

Structured clinical data on basic patient characteristics (e.g., patient name, PIN, date of birth, patient age, PSA data) are imported to ISPM Prostate from KarDa as proof-of-concept. Later, we aim to also retrieve additional well-structured data from e.g. surgery planning systems, financial records and the Electronic Medical Records (EMR) system TakeCare via KarDa. Semi-structured data, such as pathology biopsy reports, are being retrieved via the same route into ISPM for subsequent structuring. Natural language processing (NLP) algorithms specifically developed for this purpose have been implemented and evaluated. Patient-reported baseline and outcome data on general health, medication history, comorbidities, urinary and functional outcomes will be derived from structured patient questionnaires via integration to the Swedish National Prostate Cancer Register (NPCR). Selected data items collected in ISPM Prostate will be reported back to fill NPCR required forms. Data that are not registered in a semi-structured or structured way, or for other reasons not possible to access automatically from the existing IT systems, will be entered manually into ISPM Prostate. Non-structured data include e.g. EMR

reports of urological assessment, family history of (specifically prostate) cancer and comorbidities.

To enable quantification of the differences between the old clinical setting and the setting using ISPM Prostate, 12 KPIs were identified (see D3.1). Baseline data were acquired on the old setting, without the use of ISPM Prostate, and delivered in de-identified, albeit raw, format to Erasmus University. Ideally, pseudonymized data should be used for the KPIs, which requires a signed BigMedilytics Consortium Agreement and a Data Transfer Agreement between KAR and Erasmus. The CA is, at the time of writing this document, being completed for sign-off and the legal department at KAR is compiling the necessary documents for transfer of pseudonymized data.

In addition to collecting patient data in a structured way, Pilot 6 collects data relating to the postulated efficacy gain in the prostate cancer patient care flow at KAR. KPIs 1-3 and 7-9 will, in part or totally, be based on these non-clinical data (see D3.1).

Structured reporting, omitting the need for double reporting and automated output to the cancer registry should result in more efficient clinical reporting. Time-motion studies were performed to get the time currently spent – as a baseline for KPI 7 and 9 – on preparing referrals for the multidisciplinary team (MDT) conferences and post-surgery reporting. As all data are available on a single platform and presented in an integrated manner, ISPM Prostate is expected to have an impact on the quality of the MDT conference. To assess MDT quality in the current setting, the MDT Metric for the Observation of Decision-making (MDT-MODe) tool was used. Follow-up data will be collected using the same methods as baseline data, with the addition that some of the clinical and financial parameters captured in ISPM Prostate will be more easily accessed for KPI analysis.

Data Source	**Description**	**Acquisition**	**Characteristic (Size, Patients, Years, Origin/Region)**
KarDa	KAR data warehouse. Source data include EMR, surgery planning systems, lab systems etc	System integration	?
NPCR	National Prostate Cancer Registry. Quality data. PROM data	Initially: Manual entry and transfer. Later: System integration	Contains data on 98% of Swedish prostate cancer patients since the 1990s

EMR	TakeCare. Legacy system to which this project will have no automatic access.	Manual	Patient data on ~3 million SLL patients

multiple sources	**integration to data warehouse**	**data access**	**data stored in cloud**	**multi-party architecture**	**secure environment**	**transform raw / unstructured data**
yes	yes	PHI&KAR will access the data over a secure connection	no	no	yes	yes

2.3.1.2 De-Identification and anonymisation

A federated learning approach was applied to develop risk models. The approach ensures that no patient data leaves the hospital as risk models are trained and tuned on-premise and only the final risk model is shared to the external world. As such there was no need for de-identification of data. More specifically, In federated learning patient data is encrypted for protecting data privacy and security by differential privacy (Dwork, 2006) and homomorphic encryption (Gentry, 2009).

2.3.1.3 Acquisition

The various data sources will be integrated at the patient level in the tumor dashboard. Data remains on-premise and is analysed using locally installed R software and scripts. Visualizations of data distribution (e.g histograms) are generated at a population level and made available using a Shiny application over a secured https line and access is protected with a SSO.

2.3.1.4 Cleansing

Data was merged into a single unit system, e.g. for PSA values different metrics were used. Outliers were identified by visual inspection of the feature distributions, missing values were notified to the clinical/IT personnel to identify the root cause.
Feature engineering operations will be performed to create context relevant features (e.g. PSA density from PSA and prostate volume). Furthermore, feature values are validated to clinical context (e.g. PIRADS scores are always between 1 and 5, discrepancy between PIRADS scoring and biopsy outcome, etc).

Bagging and boosting techniques were applied to reduce the effect of bias on training the random forest classifier and using cross validation to estimate a fair performance over the whole dataset. An automated feature selection strategy was applied to select the optimal set of features for each prediction model.

2.3.1.5 Data Integration

An in-house developed R package is used to directly connect to the FHIR database of the ISPM prostate dashboard. The ISPM prostate dashboard is connected to the data warehouse KarDa which collects all relevant clinical data from the hospital information system. Patient record numbers and patient dates are not visible to the R user and only data population information is made visible to the user. All data is retrieved from a local server on-premise such that no individual patient information leaves the hospital.

2.3.1 Multi-velocity processing of heterogeneous data streams

Does not apply

2.3.5 Complex real-time event detection

2.3.5.1 Notifications

Table 7. Types of notifications and alerts to be issued

Need for notification services			
Notification	**Warning**	**Alarm (automated / manual reaction)**	**Other**
Likelihood score, and if available combined with a recommendation from the EAU or NCCN guidelines			-

2.3.5.2 Situations of Interest

Type of situations of interest

Table 8. Types of complex events to react on

Type of situations of interest			
Simple	**Trends (time-window / frequency based)**	**Complex (multiparameter / historical context)**	**Other**
If the value of the likelihood is above the threshold defined by the EAU or NCCN guidelines			-

2.3.5.3 Event Processing

1. **Type of event processing**

Table 9. Types of event processing actions

Filter	**Transform**	**Other**
Filter patients of which a required value is unavailable for calculating the likelihood	Percentage calculation and normalization	

2.3.5.4 Event Sources

1. **Event sources**

Table 10. Event source during complex-event processing

Stream name	Contents of stream	Stream velocity	Description of the stream
Prostate surgery decision making	Treatment decision	Weekly	Every patient is discussed during the tumour board meeting for optimal treatment decision making

2.3.5.5 Evaluation

2.4 AI Components

2.4.1 Deep learning for multilingual NLP and image analytics

2.4.1.1 Natural Language Processing

Language	How will NLP support your pilot?	How will NLP help you to reduce costs?
Swedish, Dutch, English	In comparison to structured data, text can store additional information in EHRs. For this reason, we try to integrate information from text in our prediction models. Manual extraction of data is labour-intensive. NLP will be required for acquisition of large data sets needed for high performance prediction modelling.	If including extracted information to our prediction models leads to further and meaningful improvements, this will help our pilot to reduce costs. The NLP pipeline helps to automatically extract patient data from radiology and pathology reports for use on a dashboard. The alternative would be to do this by hand (for example by a nurse) which would take a lot of time and increase cost. Now it only has to be checked, not manually extracted.

Which NLP tasks do you address?	Method	Software frameworks	Vocabularies/corpus used	Describe your method in a few sentences.	Describe your corpus/training data?
NER for disorders, findings, and anatomy in prostate biopsy pathology reports	regular expression matching	none	a few hundred pathology reports from ~10 sites in US, Netherlands, and Sweden		
item (vial) detection	regular expression matching in combination with rules	none	a few hundred pathology and radiology reports from ~10 sites in US, Netherlands, and Sweden		
measurement detection	regular expression matching	none	a few hundred pathology and radiology reports from ~10 sites in US, Netherlands, and Sweden		
negation detection	ay-like, with some improvements for our specific task	none	a few hundred pathology and radiology reports from ~10 sites in US, Netherlands, and Sweden		

2.4.1.1.1 Evaluation

2.4.1.2 Image Processing

Structured visualization of complex medical image data is a key clinical need addressed in this project. Within this project, structured visualization will have to be manually created in template schematics, but the long term goal is to automatically generate visualizations based on original data annotation performed by e.g. radiologists on source data images. This project lays the foundation for such automated image analysis.

The initial intent was to include automatic processing of image data and extracted features to include in the prediction models. This was deprioritized to favor the need of using existing risk models from clinical domains that require PIRADs scoring from radiologists and to calibrate those risk models to the local setting of the patient population at KAR.

2.4.1.2.1 Evaluation

Existing risk models from the clinical domain were in favor for the prostate surgeons as those are recommended in the EU guidelines. The existing risk models do not incorporate features that are automatically extracted from images. Therefore, automatic image analysis methodology was not a priority and is expected to be pursued in a later stage.

2.4.2 Prediction Algorithms

2.4.2.1 Task

Prediction risk of (1) pre-surgical risk of post-surgical adverse prostate cancer pathology (i.e., pathology Gleason>=7), (2) Pre-surgical risk of post-surgical advanced extent of disease (i.e., pathology disease stage >=pT3a), and (3) Pre-surgical risk of the presence of tumor infiltrated lymph nodes.

2.4.2.2 Data, Data Modelling

2.4.2.3 Features

Patient characteristics, clinical info (PSA, T stage from DRE), pathology info from biopsy (Gleason score, # of positive biopsies, etc), MRI information (PIRADS, ECE, bulging, tumor location, lymph node)

2.4.2.4 Model

The current predicted risks are detailed below, while more risk models may follow:

- Pre-surgical risk of post-surgical adverse prostate cancer pathology (i.e., pathology Gleason>=7)
- Pre-surgical risk of post-surgical advanced extent of disease (i.e., pathology disease stage >=pT3a)
- Pre-surgical risk of the presence of tumor infiltrated lymph nodes.

Models that predict risk of urinary incontinence and sexual dysfunction will follow. The output of these models is the risk to experience the relevant adverse event.
We have implemented an online learning framework to update an initial risk model with prospectively collected heterogeneous patient data. A random forest classifier was used as the prediction model in all learning strategies. Feature selection was performed based on the impact of each feature on the internally computed accuracy of the model (accuracy of model after random selection of feature at each decision tree). Note that although the RF classifier has a built-in feature selection (by prioritization), feature selection still improves the model slightly. The learning framework is connected to a clinical data dashboard which contains data elements from various medical sources in a structured way.

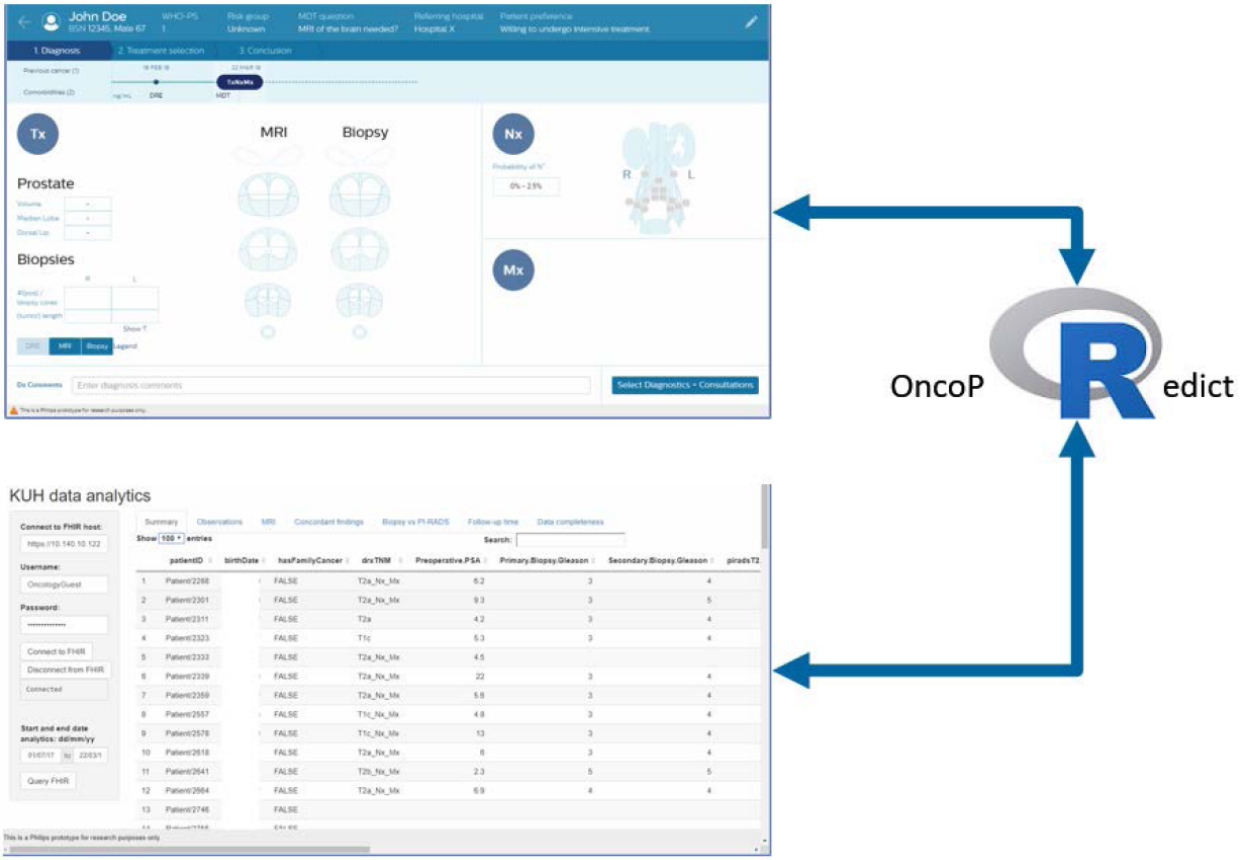

Image 15. OncoPredict

The structured data is used to execute the initially implemented risk model(s). Structured data is defined as the input variables that are available in a structured format, i.e., not in form of a variable within a medical report but in the form of a field in a database. So, the value of the variable does not need to be extracted from a report. Any prospectively collected patient data is used to update the initial model within the implemented learning framework. Over time, the initial risk model will adapt to the characteristics of the local patient population.

2.4.2.5 Evaluation

Description of Evaluation Setup (also mention if data is retrospective/prospective): The evaluation is carried out as retrospective study and evaluated within a 5 fold cross validation. In addition to that we explore how well a human doctor can solve the task of predicting one of the endpoints in the next 90 days. Then we compare this to our method and examine if a doctor can achieve better results including the prediction of our model.
Size of Data: Data includes more than 1400 patients with overall more than 100,000 data points, and then split into 70% training, 15% development and 15% test data.
Evaluation Method (e.g. ROC, F1): The model is evaluated using ROC and AUC-PC.
Results: Transplant Loss: ROC: 0.93, Rejection: 0.84, Infection: 0.80

Comparison to other related work (in terms of results): A comparison to other related work is difficult, as results might differ if cohort and data varies. This fact further highlights the needs for development of models based on local data.

2.5 Security and privacy of data access and processing

ISPM (the Philips tumour board application) is an official released product for clinical use, where all the required security measures are implemented. ISPM is installed on premise, on a windows server VM administered and controlled by KAR IT. ISPM consists of angular/typescript frontend and a java micro services backend, running on local web servers. Data is stored in a local FHIR database. To install/update software, the VM can be accessed using a secure VPN connection set up by KAR IT. ISPM runs completely locally inside the KAR firewall.

OncoPredict is a separate research prototype, also installed on the VM. It retrieves de-identified data from ISPM, to execute statistical analysis.

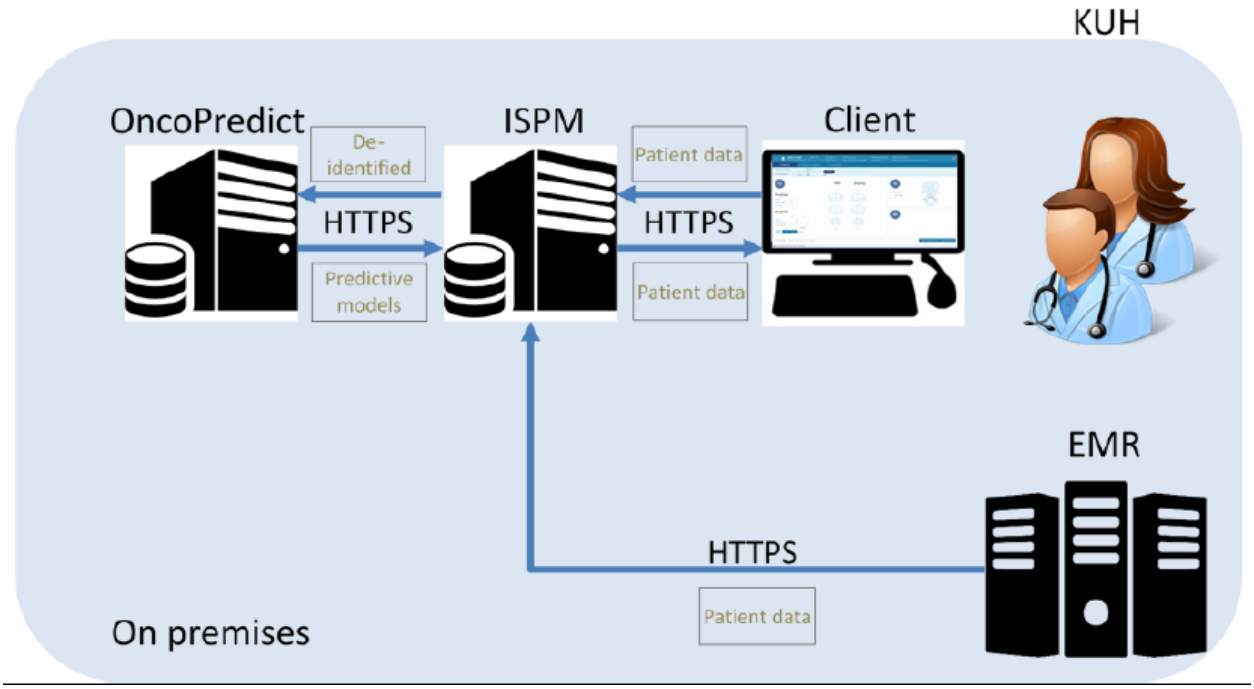

Image 16. Sw components in relation to predictive modeling and data flow

2.5.1 Access Control

2.5.1.1 Authentication

A two-factor authentication (2FA) using a token procedure is performed to gain access to ISPM from the local intranet at KAR. VPN is used to remotely connect to the VM running ISPM and

OncoPredict using Pulse Secure delivered by KAR; thereafter a Windows account login (security login).

2.5.1.2 Authorization

The current installation of the ISPM Prostate dashboard has no multiple roles defined, each user type has similar rights. Future extensions of the software include a role-based access to the ISPM dashboard where each role type is connected to a configuration of what content is to be shown to the user. Typical roles are:

- Tumour board chair
- Radiologist
- Nurse

OncoPredict enables on premise data analytics. Users can login with a 2FA with read-only access.

2.5.2 Data Protection

2.5.2.1 Data at rest

HTTP TLS encryption is used for the web interface of ISPM.

Login to the intranet of KAR via VPN. RDP connection to VM.

There is no encryption for the HD of the server. Hospital IT says: “There is no encryption for the HD of the server. Encryption is usually mostly used on discs that run the risk of getting in the wrong hands, so at our hospital they are rolling out BitLocker for laptops and stationary PC’s as they can be stolen or lost but servers are protected by perimeter security so discs are not encrypted.”

2.5.2.2 Data in transit

No data are transferred outside KAR’s intranet. All statistical analysis is performed on premises. At the end of the project all data will be transferred to KAR and the VMs are shut down and erased. During the project, which is run as an observational clinical study, ISPM will be used in parallel to legacy systems, which means that all raw data already exist in legacy systems, but not as thoroughly structured. Clinical implementation of ISPM is outside the scope of this project as mandated by Swedish laws on public procurement.

2.5.3 Auditory and logs

2.5.3.1 System Auditory

VPN logging, Windows logging by hospital IT.

2.5.3.2 Services Auditory

ISPM internal logging

2.5.4 Privacy measurements

2.5.4.1 Data Privacy Impact Assessment (DPIA)

DPIA is not required at KAR in this context.

2.5.4.2 Legal/Ethical process

The clinical study has obtained ethical permission to use clinical data without patient consent in accordance with Swedish law and GDPR with scientific research as the legal basis for processing of personal data.

2.5.4.3 Processes for complying with the current legislation

KAR is Data Controller and P is Data Processor.

2.6 Trustworthy AI

2.6.1 technology/user adoption and establishing trust

Gaining trust on the developed risk models from the users was established in several ways. Firstly, there was an external validation study performed on historic data to demonstrate the value of the risk model. Secondly, the user is shown a likelihood to an event in combination with the prediction interval to indicate a level of trust to the predicted outcome. Thirdly, the risk models are continuously updated using the local patient population and the model discriminating performance is monitored at each update cycle.

2.6.2 ethical principles

The predicted level of risk for a certain outcome (e.g. positive lymph nodes) is always shown in the context of the patient. That means that the calculated risk is shown as part of the dashboard and not in isolation. This way the clinician always has the overview and context of all the (risk) conditions of the patients. The predicted risk serves as an additional insight to support in selecting the optimal treatment strategy.
Harm is prevented by alarming the user in case that the current patient properties are outside the boundaries of the input variables of the risk models. For example, if the age of the current patient is outside the age range of the patient population used to train the risk model, the risk model will not provide a risk calculation and instead provides an appropriate message.

Bias is prevented by always using a balanced set of the patient population when training the risk model. A balanced set is created by upsampling the minority class to an equal amount of the majority class.
The random forest classifier was selected as the preferred machine learning algorithm because it includes an internal feature ranking. The ranked list of features is made available to the user as additional information to better understand what drives the algorithm most.

2.6.3 key requirements

- **Human agency and oversight**: The calculated risk scores are always shown as an additional insight to the user. The calculated risk scores are not actionable but intended to guide the clinician in deciding the optimal treatment for the patient.
- **Technical Robustness and safety**: the ISPM software is installed on premise and includes a service that continuously checks the availability of the system. Auto restarting scripts are in place in case of a sudden reboot. Concerning the machine learning algorithm, outcome data is continuously collected and used to provide a monthly update of the algorithms discriminating performances by means of an external validation study. Users are made aware that the algorithms show a poor performance.
- **Privacy and data governance**: training, testing and validating the discriminating performance of the machine learning algorithms are always performed on-premise. No data leaves the hospital.
- **Transparency**: in addition to showing the calculated risk of e.g. positive lymph nodes, the user is on request shown additional information on the model performance, the variable importance ranking and the results of the external validation using local data.
- **Diversity, non-discrimination and fairness**: the developed risk models are can be applied to all patients ; there is no restriction to any specifc patient groups or any other limitations in the clinical use as long as the patients are representative of the patient cohort that was used to develop the risk models.
- **Societal and environmental well-being**: the developed risk models will be further developed as new patient data is created during patient diagnostics. In that sense the risk models will be sustained for future patients.
- **Accountability**: The calculated risk scores for each patient are stored in logs to be able to backtrace what was shown to the user and whether that was a false negative or false positive when comparing to the later collected outcome information.

2.7 System-Interaction

2.7.1 Human-Machine Interface / GUI

ISPM is run on a web client on the KAR intranet with personal login. See screenshots under section 2.1.1. Two-factor authentication (2FA) is achieved using a token procedure (Google Authenticate) of pre-approved users. No other mobile applications are used in this pilot.

2.7.2 Education

Some very limited training of staff is needed for efficient access to and interaction with the software via the web client. Adjustment of the clinical work-process, especially in the MDT setting, has required more work and will be further evaluated and trimmed.

3. Learnings

3.1 Challenges & Barriers

- Architecture
    - Archaic and inaccessible hospital IT infrastructure
    - Ongoing development processes and recurring strategical u-turns in both K and P. As a consequence, the project is constantly aiming at moving targets.

- Processing of large structured / unstructured data sources
    - Low level of structured data sources
    - Manual work required, which includes an aspect of a pedagogical challenge, since the project aims to rather *reduce* manual double-work.

- Multi-velocity processing of heterogeneous data streams

- Complex real-time event detection
    - The implemented solution has a 24h delay in data transfer in the study setting

- Natural Language Processing
    - Swedish medical text interpretation by NLP
    - See sections 3.2 and 2.4.1 in this document and D3.3

- Image Processing
    - Differences in Swedish and international prostate sub-section definitions

- Prediction Algorithms

    - Acquisition of data
    - Strategic decisions beyond the pilot's control to change the content of solutions, preventing outcome data to be collected in ISPM as initially intended

- Security and privacy of data access and processing
    - GDPR
    - Poor institutional memory within KAR on already agreed concepts. I.e. person-depending political/strategic opinions are unstable over time ("moving-target-syndrome").

- Trustworthy AI

- System-Interaction
    - Uncertainties on how much effort should be put into integration with archaic or non-clinical systems (again "moving-target-syndrome")

3.2 Lessons Learned

- GDPR

  Shortly after the start of the BigMedilytics project the new privacy law was implemented in the EU. However, no clear guidance was given on how to interpret this law from the perspective of different types of entities (industry, academia, healthcare providers, …). This led to a very difficult situation with respect to agreeing on sharing of patient data in the context of this new legal framework.

  Given that it was known that this new law would be implemented in the EU by 2018 the parties could have been better prepared and more clear guidance could have been provided in terms of interpretation to support the legal privacy discussions within this project on patient data sharing.

- Architecture

- Processing of large structured / unstructured data sources

  Significant changes will have to be made in the way clinical data are entered by clinicians and stored in hospital IT resources to enable big-data approaches.

- Multi-velocity processing of heterogeneous data streams

- Complex real-time event detection

- Natural Language Processing

  Data extraction from clinical/medical reports via Natural Language Processing (NLP) requires a clear concept of how to deal with the extracted information. Each NLP algorithm will have a limited performance on precision and recall of certain variables. This performance may be different from variable to variable depending on how well defined (or heterogenous) the ontology of a given variable is. Consequently, the NLP

extracted data needs to be cross-checked by a medical specialist. Also, any NLP implementation requires correction mechanisms such that the specialist is able to change data that was not extracted correctly. These are all practical challenges that are important to consider and address in the context of NLP implementation. See also D3.3.

- Image Processing

- Prediction Algorithms

  The development of prediction algorithms comes with specific challenges if developed in a clinical setting where patients are stratified by other means towards certain therapeutic interventions. This will always mean that the patients that are included into the study are biased by the previous selection. In consequence only certain aspects of a risk prediction can be optimized by this approach. E.g., patients that were de-selected from an intervention by other means like imaging or an existing risk model algorithm cannot be further assessed by a new risk model that requires the outcome of said therapy. Furthermore, the access to patient data under the GDPR privacy setting needs to be considered carefully (see also above) and poses extra challenges to the development of prediction models.

- Security and privacy of data access and processing

  See above (GDPR)

- Trustworthy AI

- System-Interaction

3.3 Main (quantifiable) achievements

Major achievements of the pilot are:

- Development of a clinically useful application (ISPM - IntelliSpace Precision Medicine) to integrate heterogeneous (big!) data into a single solution to support treatment decision of primary prostate cancer during the multidisciplinary team meeting
- Implementation of the ISPM application in the form of two prototypes (the initial prototype miProstate was replaced by ISPM in 2019) into the clinical workflow at Karolinska University Hospital
- Weekly use of the ISPM application during MDT's in prostate cancer primary treatment decision making since September 2019
- Inclusion of more than 500 patients into miProstate/ISPM prototypes
- Quantification of the efficiency of the MDT in terms of i) time spent during MDT per patient, ii) presentation of the diagnostic information during MDT, iii) interaction between MDT team members
- Implementation of two clinically used prediction models and calibration of the models to the local patient population

4. Output

4.1 Papers

No publications yet. Clinical outcome is important for validation, and outcome analyses of prostate cancer care have inherent delays due to slow disease progress while long follow-up (>24 months) of side-effects related to treatment is needed for conclusiveness. Preparations are ongoing on the first manuscript, describing quality and efficiency of MDTs (see 3.3) with and without the use of ISPM in relation to radicality of treatment and patient-reported functional outcome.

4.2 Open Source & Resources (refer to ELG)

4.3 Demos

https://www.philips.com.tr/healthcare/solutions/diagnostic-informatics/oncology-informatics#triggername=close_platform