

# Development and validation of patient-level prediction models for symptoms, hospitalization and treatment initiation amongst prostate cancer patients on watchful waiting

**Sebastiaan Remmers** (✉ [s.remmers@erasmusmc.nl](mailto:s.remmers@erasmusmc.nl))

Erasmus University Medical Center Rotterdam <https://orcid.org/0000-0001-9530-6448>

**Lana Lai Yin Hui**

University of Manchester

**Carolin Schimmelpfennig**

Fraunhofer IZI

**Peter-Paul Willemsen**

University Medical Center Utrecht

**Markus Kreuz**

Fraunhofer IZI <https://orcid.org/0000-0002-0520-8391>

**Philip Cornford**

Royal Liverpool and Broadgreen University Hospitals <https://orcid.org/0000-0002-7146-9258>

**Nicolas Mottet**

University Hospital of Saint-Etienne

**Nazanin Zounemat-Kermani**

Imperial College London <https://orcid.org/0000-0003-2479-3861>

---

## Method Article

**Keywords:** big data, prostatic neoplasm, prediction

**Posted Date:** June 7th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.pex-1525/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# Abstract

The objective of this study is to develop and validate patient-level prediction models for patients on watchful waiting (WW) estimating the risk of developing symptomatic progression, hospitalization, ER visit, initiation of curative or palliative treatment, and survival. Estimation for all clinical models will be done based on 1) age and clinical measurements (e.g., PSA) 6 months before diagnosis, 2) age, clinical measurements 6 months before diagnosis, and clinical conditions one year before diagnosis. Finally, a clinically usable model will be developed based on expert clinical input. All prediction models will be implemented using Lasso logistic regression for the time at risk analyses.

## Introduction

### Introduction

Prostate cancer (PCa) is the second leading cause of cancer death in men worldwide (1). Despite high incidence rates, outcomes and survival rates for PCa have improved significantly over the years, partly due to widespread availability of prostate-specific antigen (PSA) testing (2). PSA based screening leads to earlier detection of PCa often at a stage it is amenable to treatment but years before it would have presented clinically (lead time bias) (3). However, it also leads to a significant increase in the number of men found to harbour prostate cancer the majority of whom would not have gone on to develop clinical disease (4). Historically, early detection leads to treatment much of which was unnecessary. This overtreatment, result in dysfunction of the urogenital tract including incontinence, infertility and impotence or secondary malignancies affecting patient's quality of life (5).

These two issues (overtreatment and lead time bias) have led to the development of two separate conservative approaches, active surveillance (AS) and watchful waiting (WW) for patients with PCa by the European Association of Urology (EAU) (<https://uroweb.org/guideline/prostate-cancer/>) and American Association of Urology (AUA) (<https://www.auanet.org/guidelines/prostate-cancer-clinically-localized-guideline>) guidelines. AS is an internationally accepted management strategy for men with low- and intermediate-risk PCa with a low risk of disease progression. These patients keep the option to convert to curative treatment at the time of progression (6). It attempts to reduce overtreatment by only treating those patients who are shown to progress.

Asymptomatic patients with localized disease and a life expectancy less than 10 years at time of diagnosis are not likely to benefit from radical treatment because of the lead-time bias associated with PSA testing. Instead, these patients are offered WW (i.e., symptom-guided treatment) and they may receive palliative treatment in case of progression to maintain quality of life.

Knowledge on potential risk factors for outcomes of men following WW is limited as is the natural history. The current evidence for WW mainly emerged from clinical studies which might limit the identification of potential risk factors. A European network of excellence for big data in PCa, PIONEER (<https://prostate-pioneer.eu>), partner in the Innovative Medicine Initiative's (IMI's) "Big Data for Better

Outcomes” program, aims to improve PCa care across Europe through the application of big data analytics (7). Within the context of PIONEER, the objective of the current study is to apply data driven strategies to identify predictors for symptomatic progression, hospitalization, ER visit, treatment initiation and death in order to support clinical decisions making for the management of WW.

## Objective

The objective is to develop and validate patient-level prediction models for symptomatic progression, hospitalization and palliative treatment initiation amongst prostate cancer patients on watchful waiting. In detail, we predict the 1-, 2-, and 5-year risk of developing symptomatic progression, hospitalization, ER visit, treatment initiation, and any death based on age, clinical measurements and clinical conditions to guide expectancy management for both the clinician and the patient, see Figure 1. For the model to be used in clinical practice (i.e., the clinical model), we develop a model based on expert clinical input. Discrimination will be used to compare the full “big data” model and the clinical model.

## Reagents

## Equipment

## Procedure

This study will follow a retrospective, observational, patient-level prediction design (<https://ohdsi.github.io/TheBookOfOhdsi/PatientLevelPrediction.html>). We defined the 'patient-level prediction' as a modeling process wherein an outcome is predicted within a time at risk relative to the target cohort start and/or end date. Prediction will be performed using a set of covariates derived using data prior to the start of the target cohort.

Figure 2 (from (<https://ohdsi.github.io/TheBookOfOhdsi/PatientLevelPrediction.html>)) illustrates the prediction problem we will address. Among a population at risk, we aim to predict which patients at a defined moment in time ( $t = 0$ ) will experience some outcome during a time-at-risk. Prediction is done using only information about the patients in an observation window prior to that moment in time.

We follow the PROGRESS best practice recommendations for model development and the TRIPOD guidance for transparent reporting of the model results. (8, 9). For all data sources, we refer to the appendices.

In all models we estimated the risk after 1, 2, and 5 years after diagnosis. Our population setting comprises patients with a time-at-risk window between 0 and 365 days, 0 and 730 days, and 0 and 1826 days. In all settings, the minimum lookback period applied to the target cohort is 365 days, without removing patients without time at risk or removing patients with an outcome prior to diagnosis. We included only the first exposure per patient.

## **Statistical Analysis Method(s)**

### *Algorithms*

Lasso logistic regression belongs to the family of generalized linear models, where a linear combination of the variables is learned and finally a logistic function maps the linear combination to a value between 0 and 1. The lasso regularization adds a cost based on model complexity to the objective function when training the model (10). This cost is the sum of the absolute values of the linear combination of the coefficients. The model automatically performs feature selection by minimizing this cost. We use the Cyclic coordinate descent for logistic, Poisson and survival analysis (Cyclops) package to perform large-scale regularized logistic regression: <https://github.com/OHDSI/Cyclops>

### *Model Evaluation*

Model evaluation will be based on the calibration plot and the discrimination of the internal and external validation.

### *Quality Control*

The PatientLevelPrediction package itself, as well as other OHDSI packages on which PatientLevelPrediction depends, use unit tests for validation.

### *Tools*

This study will be designed using OHDSI tools and run with R.

## *Diagnostics*

Reviewing the incidence rates of the outcomes in the target population prior to performing the analysis will allow us to assess its feasibility. The full Shiny app can be observed here: [PIONEER watchful waiting](#).

## **Data Analysis Plan**

### *Algorithm Settings*

For the time at risk analyses we use lasso regression, we use a fixed set seed and a starting lambda value of 0.01.

### *Covariate Settings*

A covariate included in the model needs to contain at least 0.001 times. In all models we specified medium term as 180 days and long term as 365 days.

### ***First model***

In the second model, we included the predictors age and all concept based clinical measurements 6 months before diagnosis defined in the OMOP Common data model.

### ***Second model***

In the final model, we extended the second model by including all concept based clinical condition one year before diagnosis and the Charlson comorbidity index.

### ***Third model***

#### *Clinical model development*

A total of six clinicians made a selection of their top 10 covariates. Consensus was reached after discussion. The included variables are:

- Grade group (levels: 1, 2, 3, 4, 5)
- PSA (levels: <10, 10-20, >20)

- Total cardiovascular disease event
- Age group (per 5 year)
- Charlson comorbidity index (levels: 0, 1,  $\geq 2$ )
- cT-stage (levels: T1, T2,  $\geq T3$ )
- Family history of PCa
- Stroke (1 year before diagnosis)
- Type 2 diabetes
- Metastatic disease extent

### *Model Development & Evaluation*

In the model development we will split the data in a train set (75%) and a test set (25%) for internal validation. The optimal lambda for the lasso regression will be assessed by 3-fold cross validation on the train set. Discriminative ability between models will be assessed by the area under the receiver operating characteristic curve (AUC). The discrimination of the clinical model will be compared against the concept-based model.

### *Strengths & Limitations*

A strength of the study is the inclusion of multiple data sources such as clinical data and claims data, all adapted with OMOP standards, allowing more generalized results. The analysis of big data may identify predictors that are currently not used in daily clinical practice. This provides a limitation but also a chance for the study. Newly identified significant predictors might not be included in clinical procedures, and therefore this study can be irrelevant for clinical questions. On the other hand, it may provide the chance to adapt current PCa treatment for the future.

A clear limitation of this study is, that in claims data the occurrence of death is not accurately presented and might be biased.

### *Protection of Human Subjects*

Local analyses were run to take into account the sensitive nature of the data. Confidentiality of patient records will be maintained always. All study reports will contain aggregate data only and will not identify

individual patients or physicians. At no time during the study will the sponsor receive patient identifying information except when it is required by regulations in case of reporting adverse events.

### *Tables & Figures*

For the incidence rate and characterization, we refer to [PIONEER watchful waiting](#).

## Troubleshooting

## Time Taken

## Anticipated Results

## References

1. Rawla P. Epidemiology of Prostate Cancer. *World J Oncol*. 2019;10(2):63-89.
2. De Angelis R, Sant M, Coleman MP, Francisci S, Baili P, Pierannunzio D, et al. Cancer survival in Europe 1999-2007 by country and age: results of EURO-CARE-5-a population-based study. *Lancet Oncol*. 2014;15(1):23-34.
3. Schröder FH, Hugosson J, Carlsson S, Tammela T, Mänttinen L, Auvinen A, et al. Screening for prostate cancer decreases the risk of developing metastatic disease: findings from the European Randomized Study of Screening for Prostate Cancer (ERSPC). *Eur Urol*. 2012;62(5):745-52.
4. Wever EM, Heijnsdijk EA, Draisma G, Bangma CH, Roobol MJ, Schröder FH, et al. Treatment of local-regional prostate cancer detected by PSA screening: benefits and harms according to prognostic factors. *Br J Cancer*. 2013;108(10):1971-7.
5. Thompson IM. Overdiagnosis and overtreatment of prostate cancer. *Am Soc Clin Oncol Educ Book*. 2012:e35-9.
6. Lam TBL, MacLennan S, Willemsse PM, Mason MD, Plass K, Shepherd R, et al. EAU-EANM-ESTRO-ESUR-SIOG Prostate Cancer Guideline Panel Consensus Statements for Deferred Treatment with Curative Intent for Localised Prostate Cancer from an International Collaborative Study (DETECTIVE Study). *Eur Urol*. 2019;76(6):790-813.

7. Omar MI, Roobol MJ, Ribal MJ, Abbott T, Agapow P-M, Araujo S, et al. Introducing PIONEER: a project to harness big data in prostate cancer research. *Nature Reviews Urology*. 2020;17(6):351-62.
8. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381-e.
9. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine*. 2015;13(1):1.
10. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996;58(1):267-88.

## Acknowledgements

Ron Herrera

Flavio Camarrone

Vasileos Sakalis

Maria Escala Garcia

Christos Chatzichristos

Billy Franks

Anke Schulz

Alex Asiimwe

Asieh Golozar

Principal investigators: Ron Hererra, Nazanin Kermani, Sebastiaan Remmers

The study is supported by the IMI2 initiative (PIONEER project).

## Figures

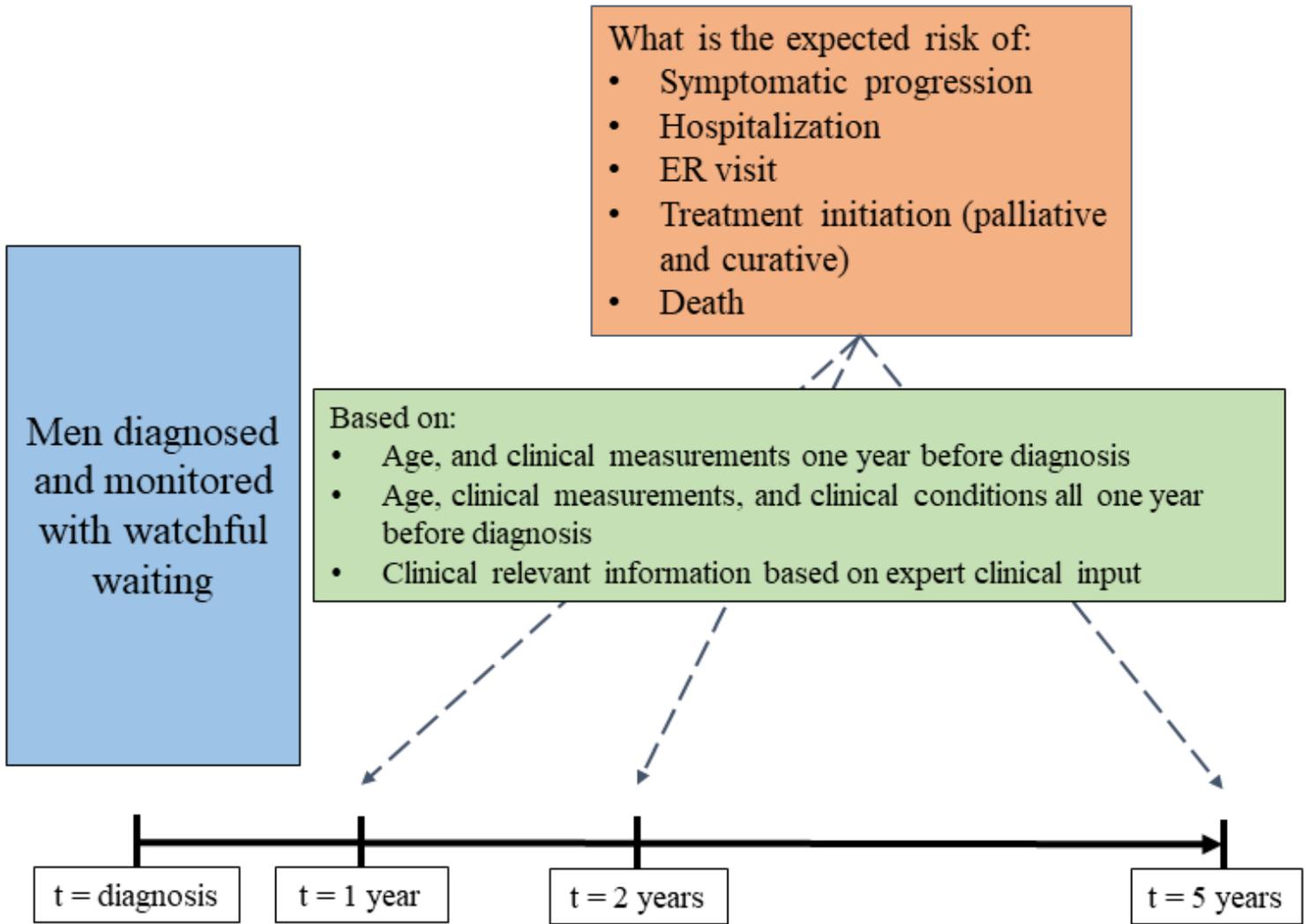


Figure 1

Overview of current study

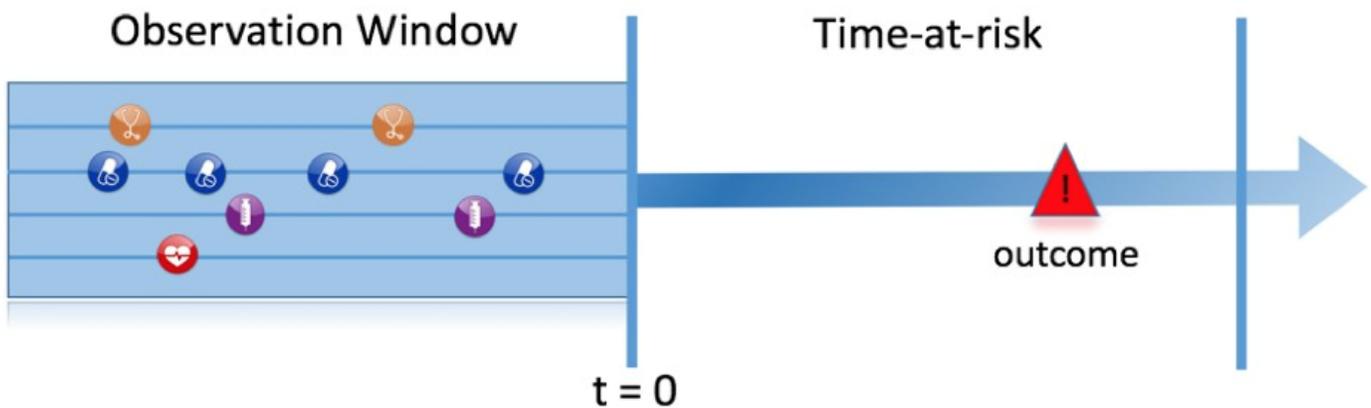


Figure 2

The prediction problem

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [FinalprotocolPackagePLPv2.docx](#)