Forecasting Population-specific 10-year Disease and Economic Burden of Prostate Cancer until 2030

SCAN-2030 Work Package 2 (WP2) – Prostate Cancer

Research Proposal

Objectives and Rationales: The objective of the study is to forecast the disease burden of prostate cancer, with the aim of helping policymakers prepare for changes in the demand for healthcare resources and treatment interventions of the population.

In light of this, the WP2 will predict the future disease burden that includes (i) incident cases, (ii) prevalent cases, and (iii) economic costs associated with prostate cancer. Using the state-of-the-art modelling technique, this study forecasts futural trends for patients with prostate cancer which ensures pre-emptive and timely planning in response to care needs.

Data source and model inputs: Results obtained from WP1 and input parameters from multiple data sources will be used to populate the models:

1. Annual disease prevalence and incidence - These parameters will be obtained from WP1 by analysing population-based cohorts identified from the routine care database.
3. To consider the potential age structure difference, age grouped population from the Census and Statistics Department of Hong Kong will be retrieved to standardize the incidence and prevalence according to age structure in 2022.
4. Cost and resource use: The cost of HRU analysed in WP1 will be utilized from records of attendance and admission, multiplied by unit costs refereeing the Hospital Authority's charge list. HRU will summarized into total annual attendance episodes or hospital days per calendrer year, under the associated cost across inpatient, outpatient and emergency settings

Statistical Analysis

We employ an AutoRegressive Integrated Moving Average (ARIMA) modeling approach. ARIMA and regression with ARIMA errors are well-established frameworks for analyzing and predicting time-series data, especially when historical patterns, autocorrelations, and potential interventions must be accounted for. For each outcome, candidate models are

automatically generated and evaluated using the auto.arima function from the R forecast package, which systematically explores a wide range of plausible ARIMA configurations. To address issues of non-normality and heteroskedasticity commonly observed in healthcare and cost data, both the raw and log-transformed versions of each outcome series are modeled in parallel.

Our modeling strategy is enhanced through the incorporation of exogenous regressors. Binary indicator variables—such as those representing the onset of the COVID-19 pandemic (dum_pan) or significant social events (dum_so)—are included as external covariates (xreg) within the ARIMA framework. These covariates are intended to capture abrupt, non-repeating shifts in the time series that are not explained by endogenous patterns alone. For the historical period (2014–2022), they reflect actual events; for future forecasts (2023–2032), the default assumption is that these events do not recur, and the covariate values are set to zero unless otherwise justified.

Model selection is guided by a combination of information criteria and error minimization. For each outcome and transformation, ARIMA models are fit to minimize both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), providing a balance between model fit and complexity. All candidate models are extracted and ranked according to their information criterion scores. When the top two models are closely matched (i.e., score gap <2), the model with the higher sum of ARIMA orders (p+d+q) is preferred, as it offers greater flexibility in capturing temporal dependencies and trends—provided that this does not introduce overfitting. If model comparison is inconclusive or model refitting fails, the best default model from auto.arima is retained.

Comprehensive model validation is performed for each selected candidate. Residuals are inspected for randomness and absence of autocorrelation, and the mean squared error (MSE) is computed on the training data. For log-transformed models, fitted values are back-transformed to the original scale prior to MSE calculation to ensure comparability across modeling strategies. The final model selection for each outcome is based on the lowest observed MSE, either on the original or back-transformed scale, thereby prioritizing predictive accuracy. Additionally, historical back-testing is conducted by comparing model predictions within the observed period to actual values, providing a practical check on the model's forecasting fidelity.

Outcome measures:

1) Annual incidence – The incident cases in a year will be forecasted based on age-standardized incidence cases from 2000 to 2020, considering the age subgroup of below or above 75 years old.

2) Annual prevalence – The annual prevalence cases, including all patients after the onset of Prostate Cancer and before death, will be employed to fit the ARIMA model and do forecasting. The age of prevalent cases will be calculated as the year difference between a specific year and the birth year, which will also be used to stratify the individuals into below and above 75-year-old sub-groups.

3) Annual and cumulative disease-related HRU costs (2021-2030) – The HRU costs for prostate cancer will include expenses related to inpatients, outpatients, and emergency department visits.

Potential pitfalls: We understand that using the length of stay as a proxy for measuring indirect costs tends to underestimate the actual expenses involved. Additionally, this represents total costs rather than costs specific to prostate cancer.