Appendix 1. Methods

Study design and data source

This is a retrospective study conducted using territory-wide electronic health records of Hong Kong. The Hospital Authority (HA) is a statutory body responsible for managing public healthcare services, including 43 public hospitals, 49 specialist outpatient clinics and 74 general outpatient clinics (GOPCs), which are mainly run by family doctors. In Hong Kong, approximately 29% of outpatient care is provided by the public sector, while over 90% of diabetes patients are estimated to have utilized diabetes care offered by the government. Patients with type 2 diabetes who are able to perform their activities of daily living independently and followed up regularly at GOPCs would be potentially invited to join diabetes mellitus complication screening (DMCS) programme. Approximately two-thirds of patients who received diabetes care at GOPCs are enrolled in the programme.[1]

The HA maintains a centralized clinical data repository that stores information on patient demographics, clinical diagnoses, prescription records, inpatient admissions, outpatient attendances, and laboratory records. Disease diagnoses were coded according to the International Classification of Disease 9th or 10th revision (ICD-9 or ICD-10), or the International Classification of Primary Care 2nd edition (ICPC-2). Data were accessed via HA Data Collaboration Lab.

Patients

Patients who were diagnosed with diabetes and received a first DMCS assessment at a GOPC between 2010 and 2019 were initially identified. Those who i) received a diagnosis of non-type 2 diabetes; ii) had a missing date of diabetes diagnosis; iii) received a diagnosis of diabetes below the age of 18 years; or iv) had a history of malignancy were excluded. Patients were followed up from a baseline DMCS assessment until a bladder cancer diagnosis, death, or December 2019, whichever came earlier. Those who i) had less than six months of follow-up; or ii) received a cancer diagnosis other than bladder cancer during follow-up were further excluded.

Covariates

Information was extracted from an initial DMCS assessment. Candidate split covariates included demographics (age and sex), duration of diabetes, medical history (chronic kidney disease, cirrhosis, ischemic

heart disease, cerebrovascular disease, heart failure, hypertension, chronic obstructive pulmonary disease, pneumonia, and family history of diabetes), medication use (anti-diabetic drugs, aspirin, non-steroidal anti-inflammatory drugs, anti-coagulants, anti-platelets, anti-hypertensive drugs, and statins), behavioral factors (alcohol use and smoking), anthropometric measurements (body mass index and waist-to-hip ratio), and laboratory measurements (HbA_{1c}, fasting glucose, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, and serum creatinine). Commonly used anti-diabetic drugs (namely metformin, sulfonylurea, insulin, and dipeptidyl peptidase-4 inhibitors) were included as candidate split covariates. Medication use was defined as whether patients had been prescribed a drug at baseline. Laboratory measurements were taken from latest results within one year of the assessment date.

Outcome

The outcome of interest was a diagnosis of bladder cancer (ICD-9: 188; ICD-10: C67) during follow-up.

Data analysis

Traditionally, Cox regression is applied to examine the association between a set of covariates and a time-to-event outcome. Nevertheless, tree-structured (or recursive partitioning) algorithms are more suitable to identify and handle potential interactions among the set of covariates on an outcome. Tree-structured algorithms are able to i) capture interaction patterns[2-3] among more than two covariates without an exhaustive search of all possible combinations among the set of covariates; ii) account for multicollinearity among the set of covariates; and iii) simultaneously account for the combined main and interaction effects of the covariates.[4] Conditional inference survival tree[5] is a tree-structured algorithm where the theory of conditional inference is embedded into its partitioning procedures. Compared to other tree-structured algorithms, it has the advantages of i) incorporating a theoretical framework; ii) minimizing bias towards covariates with many possible values; iii) avoiding overfitting; and iv) not requiring explicit pruning.

Conditional inference survival tree

At each split, a global null hypothesis of independence between any of covariates and the outcome is tested at a prespecified α-level. If rejected, a set of partial hypotheses of independence between each covariate and the outcome are tested at the same α-level. The covariate with the strongest association or minimum Bonferroni-corrected p-value is then selected as split covariate. Partitioning is recursively conducted until the global null hypothesis cannot be rejected. For continuous split covariates, the optimal cutoff value is determined by the algorithm to maximize the difference in survival outcomes between two nodes. The sequence of split covariates for each path from the root node to a terminal node represents an interaction pattern,[2] where the effects of a covariate selected at a split are conditional on covariates at its ancestor nodes. Patients were separated into mutually exclusive subgroups of most homogenous survival outcomes at terminal nodes. The α-level and maximum depth level of the tree model were set at 0.01 and 5 respectively. The cumulative incidence of identified subgroups was graphically examined. In post-hoc analyses, the associations between identified key factors and the risk of bladder cancer between comparison nodes were estimated using Cox regression and reported in adjusted hazard ratio with 95% confidence interval. Model performance was assessed using area under the curve (AUC) as metric. Data analyses were performed using R software (version 4.2.3; R Foundation for Statistical Computing, Vienna, Austria).

References

1. Lau IT. A clinical practice guideline to guide a system approach to diabetes care in Hong Kong. *Diabetes Metab J*. 2017;41(2):81-8.
2. Ramezankhani A, Tohidi M, Azizi F, Hadaegh F. Application of survival tree analysis for exploration of potential interactions between predictors of incident chronic kidney disease: a 15-year follow-up study. *J Transl Med*. 2017;15(1):240.
3. Parizadeh D, Ramezankhani A, Momenan AA, Azizi F, Hadaegh F. Exploring risk patterns for incident ischemic stroke during more than a decade of follow-up: a survival tree analysis. *Comput Methods Programs Biomed*. 2017;147:29-36.
4. Wu Z, Huang Z, Wu Y, et al. Risk stratification for mortality in cardiovascular disease survivors: a survival conditional inference tree analysis. *Nutr Metab Cardiovasc Dis*. 2021;31(2):420-8.
5. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat*. 2012;15(3):651-74.