Supplement S2: Empirical example details from Box 1

Empirical examples

Data from the Osteoarthritis Initiative (OAI) was used for two empirical examples. The OAI is a multicentre, longitudinal cohort study that included patients with (or at risk for) symptomatic femoral-tibial knee osteoarthritis (OA) with a follow-up up to 108 months, available for public access at https://data-archive.nimh.nih.gov/oai/. We extracted a large set of variables from the OAI that were measured at baseline and annual follow-up visits (12 to 108 months), these include general patients characteristics (age, gender, history of knee symptoms, physical activity, weight, care access), clinical variables (knee symptoms, radiographic signs of OA, hand OA), quality of life measurements (12-Item Short Form Survey (SF-12)), functional scores (Knee injury and Osteoarthritis Outcome Score (KOOS), Western Ontario and McMasters Osteoarthritis index (WOMAC)) and time-varying treatments (meniscectomy, knee replacement surgery, corticosteroid injections). Missing values were imputed through single imputation using predictive mean matching for continuous variables and logistic regression for categorical variables.

To investigate the impact of the different confounding adjustment methods on the outcome, two empirical examples with a time-varying treatment were selected that we previously published using data from the OAI: 1) the effect of meniscectomy (surgical removal of the meniscus) on the risk to receive knee replacement surgery and 2) the effect of intra-articular corticosteroid injections on the risk to receive knee replacement surgery.[19,20]

Statistical methods

In total, we compared nine methods that were the most commonly used adjustment methods found in the mapping review for both empirical examples: four methods that matched using baseline covariates, four time-dependent methods, and no matching. Confounding factors included in all eight correction methods were: patient characteristics (age, gender, BMI, physical activity, health care access, treatment centre, education, family history with OA, occupation), clinical variables (knee

medication use, hand OA at baseline, knee symptoms at baseline, radiographic confirmed OA), quality of life scores (SF-12 subscales), and functional scores (KOOS and WOMAC). After adjustment, Cox proportional hazard models were applied to estimate the treatment effect and confidence intervals.

The baseline methods consisted of PSM, IPW with a point treatment (yes/no), covariate adjustment using the propensity score, and conventional covariate adjustment (CCA) using baseline covariates and a point treatment. For PSM, the propensity score was calculated for every patient (the probability of a patient being assigned to the treatment given a set of observed covariates) and subsequently treated and control patients were matched using a 1:1 matching ratio without replacement, a caliper of 0.20 and a nearest neighbour matching algorithm, as nearest neighbour is commonly used and results in less biased estimates compared to the other matching algorithms.[21] Covariate balance was assessed by calculating the standardized mean difference (SMD) and by plotting the balance between patients and controls. Balance smaller or equal to 0.10 SMD were assumed to have appropriate balance.[2] IPW was performed to build a marginal structural model able to balance the covariates at baseline (marginal structural model with point treatment; patients were either labelled as treated or untreated). For IPW we used unbalanced weights and the weights were visually inspected. Similar to PSM, a 0.10 SMD was assumed to have an appropriate balance. Confidence intervals were estimated using 1000 bootstraps. Covariate adjustment using the propensity score was performed by calculating the propensity score using logistic regression and subsequently the propensity score was added to the Cox regression. Conventional covariate adjustment was performed by including the same set of covariates in the Cox regression without any prior adjustment.

The time-dependent methods consisted of time-dependent propensity score matching (tdPSM), IPW with time-varying treatment, parametric g-formula, and CCA with time-varying treatment and covariates.[5,15,17] Time-dependent propensity score matching was performed by sequentially

matching treated patients with all available controls at time of treatment using a 1:1 nearest neighbour matching algorithm without replacement using a caliper of 0.2. After matching a patient to a control, both were removed from the dataset to avoid further matches. Similar to the baseline methods, IPW was used to create a marginal structural model but with time-varying treatment and time-varying covariates. Likewise, we used unbalanced weights and the weights were visually inspected and balance was assessed. Confidence intervals were estimated using 1000 bootstraps.

Robins' g-formula (also known as parametric g-formula or parametric g-computation) is an alternative method to recover effects of time-varying treatment under untestable assumptions, given that sufficient covariates are measured to control for confounding by unmeasured risk factors.[22] The causal effect is measured by comparing the treatment effect between an always exposed- and a never exposed scenario. Conventional covariate adjustment with time-varying covariates and treatment was performed by including these variables in the Cox regression.

Finally, we performed one crude analysis by only including the time-varying treatment in the Cox regression. All analyses and simulations were performed using R (version 4.0.2, The R Foundation for Statistical Computing, Vienna, Austria) using packages 'mice', 'MatchIt', 'WeightIt', 'gfoRmula', 'plotly', 'coxphw', 'boot', and 'survival'.[12,22–29]

Results

In total, nine methods were compared for both empirical examples: four methods that adjust using baseline covariates (PSM, IPW using point treatment, CA using the propensity score, CCA), four time-dependent methods (tdPSM, IPW using time-varying treatment, parametric g-formula, CCA) and one without adjustments. (see figure in Box 1)

In the meniscectomy example, patients who underwent meniscectomy had an HR of 3.0 (95% CI: 1.97– 4.57), 2.42 (95% CI: 1.50 – 4.16), 2.41 (95% CI: 1.79 – 3.25), and 2.76 (95% CI: 2.03 – 3.76) to receive knee replacement surgery for PSM, IPW, CA using the propensity score, and CCA using the baseline covariates, respectively. The time-dependent strategies resulted in lower hazard ratios: HR of 2.00 (95% CI: 1.32 – 3.02), 2.05 (95% CI: 1.78 – 2.40), 2.03 (95% CI: 1.83 – 2.21) and 2.13 (95% CI: 1.62 – 2.79) for tdPSM, IPW, parametric G-formula and CCA, respectively. Without any adjustment, an HR of 3.15 (95% CI: 2.37 – 4.20) was found.

The results from intra-articular corticosteroid injection examples were more consistent between the baseline and time-dependent methods. Patients that receive intra-articular corticosteroid injections had a higher risk to receive knee replacement surgery with an HR of 1.64 (95% CI: 1.42 – 1.92), 1.53 (95% CI: 1.42 – 1.65), 1.58 (95% CI: 1.33 – 1.88), and 1.59 (95% CI: 1.36 – 1.87) for the baseline methods (PSM, IPW, CA using the propensity score, and CCA, respectively) and an HR of 1.61 (95% CI: 1.38 – 1.87), 1.49 (95% CI: 1.36 – 1.57), 1.65 (95% CI: 1.53 – 1.85) and 1.63 (95% CI: 1.39 – 1.91) for the time-dependent methods (tdPSM, IPW, parametric g-formula, CCA, respectively). No adjustment resulted in an HR of 2.12 (95% CI: 1.81 – 2.48).