Hybrid Matching Methods for Treatment Program Evaluation: A Case Study

Nafis Neehal*[1], Georgios Mavroudeas[1], Malik Magdon-Ismail[1], Jason Kuruzovich[1], and Kristin P. Bennett[1]

[1]Rensselaer Polytechnic Institute, Troy, NY, USA, 12180.**
{neehan,mavrog2,magdon,kuruzj,bennek}@rpi.edu

Abstract. We study a type-2 diabetes (T2D) health management program (HMP) using causal methods for treatment effect estimation on electronic health records. We use matching and survival analyses to assess T2D onset and acute care usage (emergency room or inpatient visits). To account for bias and healthcare usage changes due to the COVID-19 pandemic, we developed a hybrid matching approach that first identifies the set of potential controls based on time and other critical features and then applies matching methods. We compare results across seven state-of-the-art methods including expert-informed approaches. We find that HMP potentially improved subject health by more rapidly identifying patients with undiagnosed T2D at enrollment, allowing for timely treatment. After the initial two months, no significant differences are observed in time to T2D onset. We also found that HMP patients were less likely to seek acute care indicating improved health outcomes. We highlight practical challenges in observational health studies.

Keywords: Clinical Health Informatics, Causal Inference, Survival Analysis

1 Introduction

We use observational data to examine the effectiveness of a web-based lifestyle self-management program that has the primary goal of preventing the onset of type-2 diabetes (T2D) and improving the well-being of its clients which is being used by a midsize health payer (HP) to improve health outcomes for its members. This health management program (HMP) encourages individuals to make gradual changes and attempts to reinforce positive behavioral patterns through a 16-week program. Mobile/web applications are important to healthcare management because they offer the potential of scaling healthcare services that help individuals to implement behavioral changes, which in turn can improve health and reduce the cost of healthcare management. Most HPs who implemented such systems also want to measure the associated impact on patient outcomes. So it is important to develop methodologies to assess outcomes using possibly-biased observational data, since randomized controlled trials may not be practical.

** This work was supported by Capital District Physician's Health Plan.

Type-2 diabetes is one of the most prevalent chronic conditions, affecting over 462 million people worldwide[1]; and behavioral changes can have an important positive impact on it. A recent randomized field experiment has found that web and mobile applications for diabetes management can be effective in helping to improve patient health metrics (blood glucose and glycated hemoglobin levels) as well as reducing medical expenses and hospital visits[2]. Our goal is to help understand the degree to which this HMP can improve the health outcomes of those not yet diagnosed with T2D using historical observational data.

When evaluating the effectiveness of HMP in this study, we dealt with three major challenges: (i) T2D is a chronic disease with slow onset, so defining when T2D prevention succeeds or fails is challenging especially over the relatively short duration of this study, (ii) this is an observational study and not a randomized clinical trial, so matching methods must be used to assess any treatment effects to control for treatment selection biases, and (iii) the COVID-19 pandemic caused healthcare usage patterns to change significantly during the course of the study.

Section 3 describes how we construct the observational study to overcome these challenges: (i) To assess the treatment effects, we deploy a suite of matching methods for causal analysis ranging from propensity methods to more recent deep learning approaches which are both widely used and state-of-the-art. This ensures that our findings are robust. For our work, we adopt matching-based methods because of their popularity, effectiveness in reducing bias by balancing the feature distributions of treatment and control groups in observational studies, and finally, for its easiness in interpretability.[3] (ii) To control for confounding, we worked with experts to define variables related to the outcome. These selected variables are used to both create and assess match quality. **Select** methods match using only the expert-curated variables. The other methods use all of the available variables, either in the original (**All**) or **Latent** space. (iii) To handle the evolution of healthcare, we developed a **hybrid matching** approach. For each subject in the treatment group, we define the month of first registration as the index date. For each treatment subject, the set of possible control subjects is defined by precisely matching the index date and some of the select variables, and then propensity scoring or other matching methods are used on the remaining variables. (iv) The first outcome we analyze is the onset of Type-2 Diabetes(T2D). We restricted the analysis to patients with no prior diagnosis of diabetes at the index date and examined if the subjects were subsequently diagnosed with T2D after the index date. (v) To assess how the program affects health care usage, we choose the second outcome to be Acute Care(AC), which is defined as the length of time after the index date until a subject has an emergency visit or an inpatient visit. (vi) To capture the evolving nature of T2D and account for the right censoring of the data, we utilize survival analysis to examine HMP effectiveness.

The results in Section 4 are surprising; use of HMP increases the probability of a subject being diagnosed with T2D in the first two months after treatment. Undiagnosed T2D is harmful, while early identification and treatment of T2D have several health benefits. We also discover that the program does positively

influence the patients' well-being by reducing their emergency and inpatient visit rates. The pandemic may have exacerbated the effect as HMP patients may have been more likely to seek treatment through their primary care doctor. In Section 5 we conclude with some thoughts on the challenges in observational studies.

2 Background

Problem Formulation: For causal analysis we operate under the potential outcome framework. We utilize all the standard assumptions of causal inference[4]. Let $(X^i, T^i, Y^i) \sim P$ be our dataset where $X^i \in \mathbb{R}^d$ denotes i-th patient's d-dimensional baseline covariates, $T^i \in \{0, 1\}$ is the binary treatment variable and Y^i is the observed outcome of interest for this patient. For a treated patient the outcome is denoted by $Y^i(1)$, while a patient from the control group has the observed outcome $Y^i(0)$. In general, for evaluating efficacy of an intervention we are interested in the average treatment effect value (ATE). For randomized clinical trials, one can calculate ATE for the whole study population[5]. For observational data, this estimate may be biased since treatment subjects are not chosen at random. Thus we adopt state-of-the-art methods that match treated subjects with appropriate controls to produce unbiased ATE estimates.

2.1 Matching Algorithms

Let $P_t = \{(X^i, T^i, Y^i)|T^i = 1\}$ be the treated group and $P_c = \{(X^i, T^i, Y^i)|T^i = 0\}$ the controls. Let M be a set of matching functions we experiment with which takes the treated and control population as input and returns the matched control population. So, $X_{mc} = \forall_{f \in M},\ f(X_t, X_c)$ as we denote the matched control population as X_{mc}. For matching we deploy two popular matching functions[3] -

1. Nearest Neighbor Matching (NNM): For each treated sample X_t^i, we extract the *top-5* nearest neighbors from X_c (*top-5* defined as five samples with the lowest distance, where distance is $d = ||X_t^i - X_c^j||_2$ for any one of the top five neighbors denoted as $X_c^j \in X_c$).
2. Propensity Score Matching (PSM): Propensity Score is the probability of a patient being assigned to receive a particular treatment given a set of observed covariates. For any patient sample (X^i, T^i, Y^i), the propensity score for that patient will be $e(X^i) = Pr(T^i = 1|X^i)$. For each treated patient X_t^i, we find the *closest* patient from the controls X_c^i (*closest* defined as the control patient having the most similar propensity score to the treated sample)

We construct the matched controls (see 3.2) in multiple ways: (i) applying hybrid NNM or PSM on all original or a few selected features, and (ii) applying ML methods to generate a low-dimensional representation of the original features and then apply hybrid NNM on them. For ML methods, we use standard PCA[6], standard Autoencoder[7] and MHTM - a more sophisticated Autoencoder-like Deep Neural Network. We then compare them against the treated population and evaluate program efficacy using survival analysis.

2.2 Survival Analysis

We are interested in the time to T2D onset or Acute Care usage from the index or registration date, which is called the *survival time*. The survival function is a function that gives the probability that a patient will survive past a certain time, and this probability can be used as a proxy for survival time itself. We utilize: (i) Kaplan-Meier (KM) survival plots, (ii) Logrank tests, (iii) Cox (proportional hazards) regression, and (iv) Restricted Mean Survival Time (RMST) for our survival analysis[8,9]. These methods take into account that the data may be right-censored, e.g. if the study ended or the subjects left the insurance group.

We use Kaplan-Meier Curves to generate survival curves that represent the probability that the event (T2D or Acute Care) has not occurred after the index date at a respective time interval. The Logrank test evaluates the hypothesis that there is no difference between the populations in the survival times. We also utilize Cox's proportional hazards model to investigate the effect of several covariates on the survival time. Finally, we use RMST differences, as an alternative to Cox's Regression model for quantifying the postponement of the outcomes during a specified (restricted) interval as it corresponds to the difference between the areas under the two survival curves for the treated and control groups.

3 Methodology

3.1 Data

The proprietary observational data for this study was provided by a regional health payer organization. The data contain more than 9 million de-identified records of subjects eligible to participate in HMP, spanning from November 2017 to April 2021. For the 1,604 unique patients enrolled in HMP (treated group), we know the date of registration and their completion dates. We had records of about 350K unique patients per month to use as the unmatched control group. The records track the patient history through a number of variables which are updated monthly. Thus each patient in the data is represented with a time series of records with covariates describing their health profile. These covariates can be divided into three main parts. The first part contains 69 diagnosis/summary codes, the second part contains 3 cost-related features describing patient expenditures and the last part contains demographic and insurance information such as age, gender, and insurance type. We apply log transformations to the age and cost-associated features. To deal with potential confounding as best as possible, we use domain knowledge to identify features that are likely associated with the outcome. We narrow down the final feature list to the following: **features** (i) age (ii) total cost (iii) gender (iv) tobacco use (v) has pressure (vi) has obesity (vii) has hypertension (viii) has hypothyroid (ix) total disease count (x) acute care usage in the previous 2(ACUTE2) and 6(ACUTE6) months (xi) inpatient visits previous 6 months(ER6) (xii) emergency visits previous 6 months (IP6) (xiii) line of business (Medicaid, non-Medicaid), and denote them

as *expert features*. We include *ER6* and *IP6* since it is important to make sure that the matched patients have similar ER6 and IP6 history since the future survival of these features (combined as *Acute Care*) is studied as an outcome. We restrict the study to patients with at least 6 months of history before their registered/index date(treated/controls), since some features utilize historical information. Also, the study is restricted to patients that did not have a diabetes diagnosis at the registration/index date.

3.2 Matching Process

In an observational study, one cannot naively compare the outcomes between treated ($T = 1$) and control ($T = 0$) subjects since there are induced biases coming from the disparities in the distributions of the two groups. We develop and deploy a hybrid matching scheme and we explore several matching algorithms as a part of this hybrid matching process to produce controls similar to the treated population. The methods are from two broad categories –
(i) **Propensity Score Matching (PSM) based methods:** PSM Select, PSM All (ii) **Nearest Neighbor Matching (NNM) based methods:** NNM Select, NNM All, Principal Component Analysis (*PCA Latent*), Autoencoder (*AE Latent*) and Member Health Trajectory Model (*MHTM Latent*). In PSM Select and NNM Select, we run the PSM and NNM algorithm on only the *expert features*, while in PSM All and NNM All, we run the PSM and NNM algorithm on all the features. In PSM Latent, AE Latent, and MHTM Latent, all of the features are used to create a latent space, and NNM is done in latent space. In all cases, the time frame of treated and controls was matched as a pre-filtering step. As subjects enroll in HMP at different times, so it is very important to match on the enrollment date to control for changes in healthcare and different follow-up times.
NNM Hybrid Matching Process: For all five NNM based methods, we apply this additional pre-filtering step along with the NNM algorithm. Hence, we denote this matching scheme as *hybrid matching*. It combines K nearest neighbors (KNN), with exact and coarsened exact matching[10]. For all methods, we always match the month as part of the multi-stage pre-filtering step. We denote a treated subject i with registration date at time t_i as X_t^i and a control subject j at time t_j as X_c^j. Simultaneously, we define the corresponding data including all the subjects in the two groups as X_t, X_c for the treated and controls respectively. After the matching is done we acquire the matched control group X_{mc}. Algorithm 1 describes the hybrid matching process in detail.

After filtering the controls in this manner, we further define three types of features $X_{ex}, X_{int}, X_{nn} \in X$ depending on how we apply the matching algorithm on them, where X is all baseline covariates. For X_{ex} features we match exactly on the values between the treated and control subjects. For X_{int} we try to match around an interval of the treated subjects' values; for each feature $X_{int}^i \in X_{int}$ we define a list with the allowed intervals to match, $H = \{h_i \forall X_{int}^i \in X_{int}\}$. Finally, we match based on the Euclidean distance between the treated and the control subjects using only the X_{nn} features.

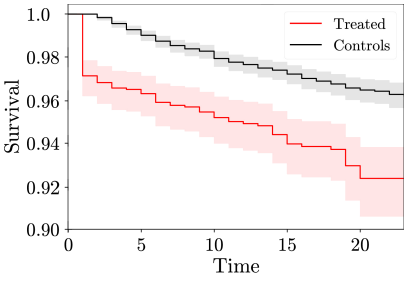
Algorithm 1: Hybrid NNM Matching of treated and control subjects.

Input: : $X_t, X_c, K, X_{ex}, X_{int}, X_{nn}, H$
Output: Matched controls X_{mc}
$X_{mc} = \{\}$;
for *patient i in* X_t *with registration time* t_i **do**
  X_{c1}: filter X_c to extract controls only with time t_i;
  X_{c2}: $\forall\, X_c^i \in X_{c1}$, filter X_{c1} s.t $X_t^i(X_{ex}) = X_c^i(X_{ex})$;
  X_{c3}: $\forall\, X_c^i \in X_{c2}, \forall\, X_{int^j} \in X_{int}$, filter X_{c2} s.t $X_t^i(X_{int^j}) \in [X_c^i(X_{int^j}) \pm h_j]$;
  X_{mc}^i: find the K nearest neighbors of X_t^i from X_{c3} based on the Euclidean distance of X_{nn} features;
  $X_{mc} = \{X_{mc}; X_{mc}^i\}$;
end

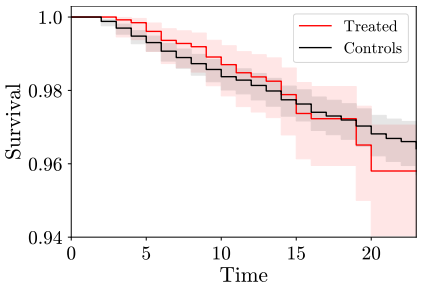
Latent Space Methods: Using machine learning methods for matching on a latent subspace to generate the matched controls has become increasingly popular[11,12]. Thus, we also use one linear and two nonlinear latent space matching methods (n=16 components): PCA Latent, Autoencoder (AE Latent), and Member Health Trajectory Model (MHTM Latent). AE Latent and MHTM Latent both models have almost identical architecture – an Input Layer, 5 consecutive hidden blocks (each block contains a Dense Layer with ReLU activation, followed by a dropout Layer with p=0.5 dropout probability) of 64, 32, 16, 32 and 64 dimensions respectively; with the 16-dimensional layer being the representation layer. Both input and output layers have the same number of nodes. For all three methods, the latent features represent a summary of all the input features over the last year. We take the means of the monthly records as the aggregated patient representation used for training. The AE Latent model attempts to reconstruct input values in output while the MHTM Latent model outputs *next* year's aggregated patient profile. We apply this step to deal with the seasonality in data and hope to appropriately capture the significant underlying relationship amongst the original features in the representation layer of these models. These latent models were created for another project, and we include them to provide a robust set of results. We leave a more intensive investigation and comparison of input representation and latent space methods as future work.

4 Results and Discussion

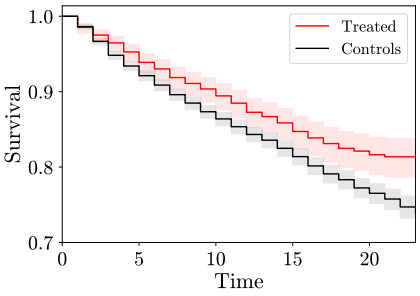
Survival curves for the NNM Select for multiple outcomes and logrank p-values for all methods are shown in Fig. 1. These p-values define how similar the treated and the control populations are in terms of survival probabilities. In Table 1, we present the RMST difference values measured after each 6 month period from the treatment enrollment for up to 18 months. A negative RMST difference value indicates the treated population tends to experience the outcome sooner than the controls and a positive RMST difference value indicates the exact opposite.



(a) T2D from idx date (p<0.05)



(b) T2D from idx date+2 months (p=0.96)



(c) Acute Care from idx date (p<0.05)

Method	T2D	T2D+2	Acute
NNM Select	**	0.96	**
PSM Select	**	0.71	**
NNM All	**	0.65	0.10
PSM All	**	0.78	**
PCA Latent	**	0.94	**
MHTM Latent	**	0.39	**
AE Latent	**	0.37	0.05

(d) Logrank p-values. ** means p < 0.05

Fig. 1: (a-c) PSM Select Kaplan-Meier Curves showing PSM Select treated (red) and control (black) survival cuves and (d) Logrank p-values for all methods.

Type-2 Diabetes Diagnosis: The Kaplan-Meier survival curves for NNM Select shown in Fig. 1(a) demonstrate that the HMP patients (red curve) are diagnosed with T2D faster and at a higher rate than controls (black) with p-value < 0.05 by the logrank test. The logrank p-values show that this finding is consistent for all the methods. We see that the drop primarily occurs in the first few months. Hence, we also examine the T2D survival probabilities after removing all patients having T2D within the first 2 months after the index date (see Fig 1(b) and p-values in Fig 1(d).) We observe, that after the first two months, there is no significant difference between the time to the first T2D diagnosis.

Recall that T2D is a chronic disease that develops over time and that all patients analyzed were not formally diagnosed with T2D at the index date. We hypothesize that HMP helped patients with undiagnosed T2D seek evaluations by primary care providers that led to a T2D diagnosis. Also, patients may have chosen to participate or be recommended to the program because they were at high risk for T2D. Timely diagnosis and treatment of T2D can improve patient outcomes. HMP's effect on earlier diagnosis may have been especially valuable during COVID-19 when many patients were not seeking routine care. A randomized study or further investigation of patients with fast onset T2D diagnoses would be needed to fully understand this effect.

Table 1: RMST Analysis for T2D and Acute Care after Index Date

	Diabetes Diagnosis			Acute Care		
	6 Months	12 Months	18 Months	6 Months	12 Months	18 Months
NNM Select	-0.147	-0.315	-0.497	0.069	0.241	0.441
PSM Select	-0.148	-0.319	-0.506	0.060	0.222	0.428
NNM All	-0.140	-0.297	-0.469	0.028	0.135	0.257
PSM All	-0.144	-0.312	-0.496	0.094	0.396	0.791
PCA Latent	-0.141	-0.291	-0.469	0.054	0.191	0.356
MHTM Latent	-0.148	-0.325	-0.518	0.048	0.176	0.379
AE Latent	-0.148	-0.331	-0.536	0.047	0.158	0.301

Acute Care: After the surprising results on the onset of T2D, might HPM contribute to patient health in other ways too? One possibility is to examine healthcare expenditures, but the pandemic changed healthcare usage. We thought acute care conditioned on time would be a more consistent measure of health outcomes. In Fig. 1(c), we present Kaplan-Meier survival curves for time to obtaining acute care after index date for NNM Select, and Fig. 1(d) presents all the logrank p-values for all the methods. The RMST values for all methods are presented in Table 1. Note that the curves for treated and control are almost identical for the first two months so we did not do any special time analyses. We observe that for all methods, the treated population has a much lower probability of obtaining acute care than the controls, and the effect is significant except for AE Latent and NNM All. We also perform a Cox's proportional hazard treatment analysis based on NNM Select matching controlling for all the *expert features* and the treatment. We get a treatment coefficient of -0.271 (p-value <0.05) indicating that HMP patients seek less acute care.

Comparison of Matching Methods: Fig. 1 and Table 1 show that all methods produce the same conclusions that HMP patients are diagnosed with T2D faster and are less likely to use acute care with the minor exception that the logrank p-values for AE Latent and NNM All are slightly above 0.05. We note, however, that the RMST analysis in Table 1 indicates that the magnitude of the effect produced by the 7 methods varies. NNM Select and PSM Select benefit from domain knowledge, thus are our best estimate of the treatment effects. They produce almost identical results. NNM and PSM using all the features produce very different results at 12 and 18 months indicating that they match on different controls. The latent space methods use hybrid NNM in the latent space. They achieve results much closer to NNM Select and PSM Select than NNM All and PSM All. Although they were not trained specifically for this problem, they are trained using the average features averaged over the prior 12 months for all patient data with at least 10 months of data in the last year. Yet they produce similar results without the use of domain knowledge except for within hybrid matching. We also examine the quality of the matches. For the selected features, Table 2 compares the means (with p-value) of the treated and matched controls

Table 2: Comparison of Treated and Control Means (*p<0.05) for All Methods.

	Treated	Matched Controls							All Controls
Features		NNM Select	PSM Select	NNM All	PSM All	PCA Latent	MHTM Latent	AE Latent	
Age	50.74	50.82	49.71*	50.80	50.81	50.77	50.78	50.79	52.64*
Total Cost	712.1	641.0	634.8	589.3*	708.0	765.6	749.3	827.0	899.38
Gender	0.21	0.21	0.22	0.21	0.27*	0.21	0.21	0.21	0.43*
Tobacco	0.06	0.05	0.06	0.07	0.10*	0.09*	0.10*	0.09*	0.11*
Pressure	0.00	0.00	0.01	0.00	0.01*	0.00	0.00	0.00	0.02*
Obesity	0.50	0.49	0.50	0.50	0.32*	0.30*	0.29*	0.29*	0.30*
Hypertension	0.34	0.33	0.32	0.35	0.32	0.25*	0.25*	0.25*	0.38*
Hypothyroid	0.10	0.08	0.08*	0.09	0.09	0.09	0.09	0.08*	0.09
Disease Count	2.91	2.87	2.66*	2.73*	2.82	2.42*	2.40*	2.38*	3.36*
Acute Care (Prior 2 Mon.)	0.04	0.03	0.02*	0.03	0.03	0.03	0.03	0.03	0.06*
Acute Care (Prior 6 Mon.)	0.12	0.11	0.08*	0.11	0.11	0.11	0.11	0.11	0.17*
Inpatient (Prior 6 Mon.)	0.03	0.03	0.02	0.03	0.03	0.03	0.03	0.03	0.06*
ER Visits (Prior 6 Mon.)	0.09	0.09	0.06*	0.09	0.08	0.09	0.09	0.09	0.12*
Business Line	0.96	0.96	0.90*	0.96	0.82*	0.96	0.96	0.96	0.82*

for each method. The means for all eligible controls are significantly different for 12 of the 14 variables. There are no statistical differences in the means for NNM Select and the treated population. All methods except PSM Select match all of the variables related to prior acute care. NNM All finds patients that are significantly healthier as indicated by lower Total Costs and Disease Counts. PSM All has different distributions for several categorical variables.

5 Conclusions

This case study illustrates the practical challenges of evaluating the effectiveness of HMPs using observation studies based on electronic health records. Our strategy is to examine multiple outcomes using a suite of 7 different matching methods including classical ones and deep learning-based advanced ones. We start with 77 features and then use expert advice to create an expert-curated set of features known to be relevant to the outcome. We match by month for

all methods to control for the changing health care conditions and seasonality. For NNM, we match precisely on features important for the problem and applied the nearest neighbor matching to the rest. The results are very consistent across all the methods. We find that HMP increased the likelihood of a new T2D diagnosis in the first two months but found no significant difference after that, and patients in the HMP were less likely to need inpatient or emergency care for a greater time. Evaluating the match quality, we find that the methods matching on selected features performed the best. The latent space results shown here are promising but many modeling improvements are possible. One limitation of this work is that our models do not consider how much of HMP the subjects completed. In the future, semantically-aware tools based on latent space models trained on large EMR datasets could be used to more effectively and efficiently evaluate many different treatments for both program-specific and general outcomes leading to better and cost-effective health care.

References

1. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of Type 2 Diabetes–Global Burden of Disease and Forecasted Trends. Journal of Epidemiology and Global Health. 2020;10(1):107.
2. Ghose A, Guo X, Li B, Dang Y. Empowering Patients Using Smart Mobile Health Platforms: Evidence from a Randomized Field Experiment. Forthcoming at MIS Quarterly. 2021.
3. Stuart EA. Matching Methods for Causal Inference: a Review and a Look Forward. Statistical Science: a Review Journal of the Institute of Mathematical Statistics. 2010;25(1):1.
4. Rubin DB. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. Journal of the American Statistical Association. 2005;100(469):322-31.
5. Holland PW. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945-60.
6. Abdi H, Williams LJ. Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(4):433-59.
7. Charte D, Charte F, Del Jesus MJ, Herrera F. An Analysis on the Use of Autoencoders for Representation Learning: Fundamentals, Learning Task Case Studies, Explainability and Challenges. Neurocomputing. 2020;404:93-107.
8. Efron B. Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve. Journal of the American Statistical Association. 1988;83(402):414-25.
9. Royston P, Parmar MK. The Use of Restricted Mean Survival Time to Estimate the Treatment Effect in Randomized Clinical Trials when the Proportional Hazards Assumption is in Doubt. Statistics in Medicine. 2011;30(19):2409-21.
10. Iacus SM, King G, Porro G. Causal Inference without Balance Checking: Coarsened Exact Matching. Political Analysis. 2012;20(1):1-24.
11. Chu Z, Rathbun SL, Li S. Matching in Selective and Balanced Representation Space for Treatment Effects Estimation. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management; 2020. p. 205-14.
12. Johansson F, Shalit U, Sontag D. Learning Representations for Counterfactual Inference. In: International Conference on Machine Learning. PMLR; 2016. p. 3020-9.