Manuscript 4

SuperLearner

Title:

Marginal structural models using calibrated weights with SuperLearner: application to type II diabetes cohort

Authors:

Sumeet Kalia [1] ; Olli Saarela [1]; Michael Escobar [1]; Tao Chen [2]; Braden O'Neill [1,5] ; Christopher Meaney [1]; Jessica Gronsbell [1]; Babak Aliarzadeh [1]; Ervin Sejdić [2]; Rahim Moineddin [1,3]; Conrad Pow [2]; Frank Sullivan [4]; Michelle Greiver [1,2]

Affiliations:

[1] University of Toronto; [2] North York General Hospital; [3] Institute of Clinical and Evaluative Sciences; [4] University of St Andrews; [5] Unity Health Toronto

Contents

4.1 Abstract

Although machine learning has permeated many disciplines, the convergence of causal methods and machine learning remains sparse in the existing literature. Our aim was to formulate a marginal structural model for estimating diabetes care provisions in which we envisioned hypothetical (i.e. counterfactual) dynamic treatment regimes using a combination of drug therapies to manage diabetes: metformin, sulfonylurea and SGLT-2i. The binary outcome of diabetes care provisions was defined using a composite measure of chronic disease prevention and screening elements (Nietert et al., 2007) including (i) primary care visit, (ii) blood pressure, (iii) weight, (iv) hemoglobin A1c, (v) lipid, (vi) ACR, (vii) eGFR and (viii) statin medication. We used several statistical learning algorithms to describe putative causal relationships between the prescription of three common classes of diabetes medications and quality of diabetes care using the electronic health records contained in National Diabetes Repository. In particular, we generated an ensemble of statistical learning algorithms using the SuperLearner framework based on the following base learners: (i) least absolute shrinkage and selection operator, (ii) ridge regression, (iii) elastic net, (iv) random forest, (v) gradient boosting machines, (vi) neural network. Each statistical learning algorithm was fitted using the pseudo-population generated from the marginalization of the time-dependent confounding process. The covariate balance was assessed using the longitudinal (i.e. cumulative-time product) stabilized weights with calibrated restrictions. Our results indicated that the treatment drop-in cohorts (with respect to metformin, sulfonylurea and SGLT-2i) may improve diabetes care provisions in relation to treatment naïve (i.e. no treatment) cohort. As a clinical utility, we hope that this article will facilitate discussions around the prevention of adverse chronic outcomes associated with diabetes through the improvement of diabetes care provisions in primary care.

4.2 Keywords

Longitudinal interventions; Machine Learning; SuperLearner; Electronic Health Records; Primary Care; Chronic Disease Prevention, Screening and Management;

4.3 Introduction

We may describe the multi-faceted data analytics landscape using three paradigms: (i) data exploration, (ii) data inference and (iii) data prediction. To date, most causal methods focused on a data inference paradigm in which hypothetical interventions are constructed, and the philosophical discussions around *"causal methods"* can be traced back many centuries (Hume, 1739). Although the methodological techniques for the validation of *"causal prediction models"* are still in their infancy (Lin et al., 2021), our aim was to formulate marginal structural models in which we envisioned hypothetical (i.e. counterfactual) treatment regimes. We construct the hypothetical treatment cohorts using the referent modality (i.e. treatment naïve cohort) and indexed modality (i.e. treatment drop-in cohort). We may describe the "treatment-naïve" cohort as the absence of treatment regimen while the *"treatment drop-in"* cohort as the initiation of treatment post-baseline (Lin et al., 2021). For example, we may consider a hypothetical contrast in which the patients are not prescribed glucose-lowering medications during the study period and we may use this cohort to describe the referent regimen in relation to treatment regimens imposed using the treatment drop-in cohorts. We define the hypothetical interventions for diabetes care provisions using a combination of glucose-lowering medications including metformin, sulfonylurea and sodium-glucose co-transporter-2 inhibitors (SGLT-2i) (Greiver et al., 2021).

It is essential to distinguish between the etiological and the intervening genres of causality in medicine (Karp and Miettinen, 2014). In this article, we like to emphasize that the hypothetical treatment modalities of glucose-lowering medications were not assumed to be etiological with respect to the diabetes care provisions. Rather the focus was limited to the estimation of diabetes care provisions in which we deliberately intervene on longitudinal treatment regimes indexed with respect to annual calendar time. There is an emerging focus in causal literature around precision medicine with individualized treatment regimes (Shalit, 2020). We characterized the individual-level treatment regimes with respect to the clinical profile of each patient using the conditional average treatment effect. In particular, we described the clinical profile of each patient presenting at primary care clinics within a calendar year using the time-varying outcome-predictors (i.e. effect modifiers) including annual laboratory requisitions (e.g. hemoglobin A1c), vaccination (e.g. influenza), lifestyle information (e.g. smoking documentation), diagnostic codes and billing codes. Although the marginal structural model supported the individualized estimation, we chose to simplify the causal risk difference to population-averaged estimation as the validity of individualized treatment regimes in causal literature is often debated (Rose and Rizopoulos, 2020).

In principle, we may create a clinical model with respect to hypothetical interventions using two methodologies: (i) Bayesian construct in which posterior inference is derived using prior information gathered from previous studies, (ii) frequentist construct in which inference is derived from the data using the likelihood-based contributions. For example, in the context of individualized Bayesian prediction, Alaa and van der Schaar (2017) used *"precision-in-estimating heterogeneous effects"* as the loss function to minimize the error between factual outcomes and posterior counterfactual variance while Arjas (2014) used Bayesian non-parametric formulation with marked point process to predict the outcome with respect to counterfactual intervention assuming continuous-time. Unlike the earlier work, this article focuses on discrete time-intervals in which the hypothetical models are formulated in the presence of time-dependent treatment and treatment-confounder feedback (Hernán and Robins, 2022). We apply the frequentist construct in which the hypothetical interventions with respect to appropriate treatment modalities are conceptualized to estimate the diabetes care provisions in next calendar year. We assume that the

individuals are independent and identically distributed (i.i.d.) whereby we do not have the necessary information to distinguish between two pairs of individuals (De Finetti, 1974). The i.i.d. assumption allow us to characterize the discrete time processes using the non-distinguishable individual-level indices in frequentist construct.

In recent literature, the unification of machine learning with inference has been the dominant subject in which semi-parametric theory is used to create machine learning based effect measures, augmented together with causal identifiability assumptions, to produce causal estimates (Rose and Rizopoulos, 2020). However, less emphasis is placed on causal estimation task using deep learning in which more abstract representations can be computed in relation to the less abstract ones (Goodfellow et al., 2016). There is a growing recognition and understanding of structural racism when machine learning algorithms are implemented, especially in the context of many health and judicial applications (Mehrabi et al., 2021; Robinson et al., 2020). With this in mind, our aim is to deploy the hypothetical estimation of diabetes care provisions (in future) with the emphasis on assessing and eliminating bias arising due to temporal confounding and other epidemiological sources. For example, the use of older glucose-lowering medications (e.g. Sulfonylurea) might be associated with worse health outcomes than newer glucose-lowering medications (e.g. SGLT-2i). We may describe this phenomena as "confounding by indication", and this phenomena coupled with unmeasured or hidden confounders may thwart our ability to correctly identify the causal estimates (Shalit, 2020). Although the randomization procedure in controlled experiments nullifies these causal challenges whereby the controlled experiments are by design unconfounded and associations imply causations (Hernán and Robins, 2022), we need to account for these causal and statistical challenges when drawing unconfounded estimation from longitudinal cohorts with observatinal design. This, in turn, allow us to generate reliable estimation with greater scope of generalizability when the application of machine learning algorithms was shifted from training sample to test or validation sample.

4.3.1 Motivation and Knowledge gap

Although machine learning has permeated many disciplines, the convergence of causal methods and machine learning remains sparse in the existing literature (Rose and Rizopoulos, 2020). The objective of this article was to demonstrate the application of SuperLearner using the amalgamation of the machine learning algorithms in the context of hypothetical interventions for diabetes care provisions using the primary care electronic health records (EHRs). Although the hypothetical interventions are not directly observable in practical sense, the aim of this study is to facilitate the discussion around the prevention of chronic adverse outcomes associated with diabetes through the improvement of diabetes care provision in primary care.

4.4 Materials and Methods

The material section describes the data source, and the methods section is split into two sub-sections: (i) notational framework and (ii) machine learning algorithms. The notational framework describes the causal notation, followed with identifiability assumptions and the stabilizing weight function to account for time-dependent confounding process. A collection of diverse machine learning algorithms are described so that we can construct the stacked estimation using the SuperLearner framework.

4.4.1 Data Source

Diabetes Action Canada's National Diabetes Repository (NDR) was created in 2017 with the collective goal of enhancing care among patients with diabetes. The NDR collates electronic health records (EHRs) on patients living with diabetes across multiple practice-based research networks (PBRNs) located in Alberta, Manitoba, Quebec, Ontario, and Newfoundland. As of 2020Q2, the NDR collects information on 148,707 diabetes patients distributed across 1,342 primary care providers with 145,558 age and sex matched controls (i.e. patients not living with diabetes) for comparative research. The EHRs in NDR contain patient-level demographics, medical diagnosis, procedures, medications, immunization, laboratory test results, vital signs and risk factors. Since the EHRs in NDR comprises of PBRNs across multiple provinces in Canada, we limited the scope of the data source for this study to PBRNs within Ontario: (i) University of Toronto Practice-Based Research Network (UTOPIAN), (ii) Eastern Ontario Network (EON). This allowed us to control for the possibility of data heterogeneity arising due to uncontrollable sources (e.g. data extraction practice; commercialized software of EHR systems; provincial health regulatory bodies) in EHRs (Shi et al., 2020b). The estimation tools developed using the causal methods were more likely to be generalizable and portable when applied to homogeneous EHR data sources, as the possibility of distributional shift of the training set was reduced (Amodei et al., 2016). Analyses were performed using the R software (v.4.1.0) in Secure Analytic Virtual Environment at the Centre for Advanced Computing located at Queen's University.

4.4.2 Notational framework

We specify the notational framework using the potential outcomes (i.e. counterfactual outcomes). At first, we introduce the notation for longitudinal repeated-measures outcomes, followed by sequential variants of identifability assumptions. We formulate a stabilizing weight function with calibrated restrictions to account for time-dependent confounding process.

Notation

A longitudinal model is considered for n individuals ($i = 1, ..., n$) in j discretized calendar time points (i.e. $j = \{2016, 2017, 2018, 2019\}$). We denote the longitudinal binary outcome of diabetes care provisions as Y_{ij}. The treatment at time t with respect to the eight combinations of glucose lowering medications (i.e. metformin, sulfonylurea, SGLT-2i) is denoted as A_{ij}. We denote the patient demographics with respect to k^{th} baseline covariates as X_{ik}. We partition the time-varying covariates as confounders (i.e. common cause of treatment process and outcome process) and outcome-predictors (i.e. effect-modifiers). The time-varying covariates include International Classification of Disease version 9 (ICD9) codes contained in cumulative patient profile (CPP), and Anatomical Therapeutic Chemical Classification System (ATC) medications codes while time-varying outcome-predictors include vaccination, lifestyle information, annual laboratory requisition, billing ICD9 codes and Ontario Health Insurance (OHIP) billing codes. We denote the time-varying covariates as L_{ijk} and time-varying outcome-predictors as M_{ijk} for k^{th} predictors of i^{th} individual belonging to j^{th} calendar year. We construct the histories with respect to discrete time points for treatment as $\bar{A}_{ij} = \{A_{i1}, A_{i2}, ...A_{ij}\}$, time-varying covariates as $\bar{L}_{ijk} = \{L_{i1k}, L_{i2k}, ...L_{ijk}\}$, time-varying outcome-predictors $\bar{M}_{ijk} = \{M_{i1k}, M_{i2k}, ...M_{ijk}\}$, and repeated-measures outcomes as $\bar{Y}_{ij} = \{Y_{i1}, Y_{i2}, ...Y_{ij}\}$. For the sake of brevity, we suppress the index for individual i in some instances with the assumption that the random vector for each individual i is

sampled independently with respect to other individuals.

Diabetes care provision

We describe *"diabetes care provisions"* using a modification of the summary quality index inspired by Grunfeld et al. (2013) and Nietert et al. (2007). We define the longitudinal primary endpoint for diabetes care provisions as the sum of eight elements as

$$\begin{aligned}(\text{Diabetes care elements})_j = \mathbb{1}(\text{Visit count} \geq 2)_j &+ \mathbb{1}(\text{Blood pressure count} \geq 2)_j \\ &+ \mathbb{1}(\text{Weight count} \geq 2)_j \\ &+ \mathbb{1}(\text{Hemoglobin A1c count} \geq 2)_j \\ &+ \mathbb{1}(\text{Lipid count} \geq 1)_j \\ &+ \mathbb{1}(\text{ACR count} \geq 1)_j \\ &+ \mathbb{1}(\text{eGFR count} \geq 1)_j \\ &+ \mathbb{1}(\text{Statin count} \geq 1)_j\end{aligned}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function indexed with respect to calendar year j. We further define a composite binary endpoint using the sum of eight elements of diabetes care provisions within a calendar year: (i) primary care visit, (ii) blood pressure, (iii) weight, (iv) hemoglobin A1c, (v) lipid, (vi) albumin to creatinine ratio (ACR), (vii) estimated glomerular filtration rate (eGFR) and (viii) statin medication. We binarize the longitudinal score of $(\text{Diabetes care elements})_j$ as

$$Y_{ij} = \begin{cases} 1 = \text{Adequate to optimal service when } (\text{Diabetes care elements})_j \in \{4,5,6,7,8\} \\ 0 = \text{Less than adequate service when } (\text{Diabetes care elements})_j \in \{0,1,2,3\} \end{cases} . \quad (4.1)$$

Identifiability assumptions

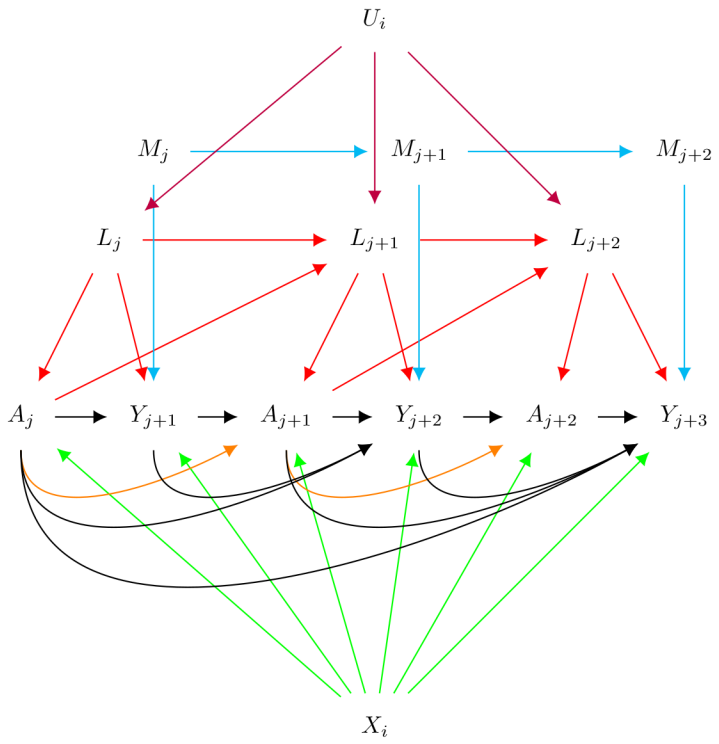
Identifiability assumptions are necessary to ensure that we can estimate the causal estimands from longitudinal studies with observational design. The necessary identifiability assumptions include: (i) sequential exchangeability; (ii) sequential postivity; (iii) sequential consistency (Hernán and Robins, 2022). We may describe the sequential exchangability as "*no unmeasured confounding*" whereby the probability of treatment assignment at each discretized time point j is independent of the potential outcome (with respect to the causal treatment regimes) conditioned on the observed history. We may succinctly write the sequential exchangability assumption as $Y_j^g \perp\!\!\!\perp A_j | \bar{\mathcal{H}}_{j-1}$ where Y_j^g denotes the potential outcome under the causal treatment regime g, and where $\bar{\mathcal{H}}_{j-1} \equiv \{\bar{A}_{i,j-1}, \bar{L}_{i,j-1,k}, \bar{M}_{i,j-1,k}, \bar{Y}_{i,j-1}\}$ is the observed history up to and including time point $j-1$. We may describe the sequential positivity assumption as the non-zero probability of treatment assignment at each time point j conditional on the observed history $\bar{\mathcal{H}}_{j-1}$. We may succinctly write the sequential positivity assumption as $P(A_j | \bar{\mathcal{H}}_{j-1}) > 0$. The sequential consistency assumption is used to connect the potential (i.e. counterfactual) outcome with respect to the causal treatment regimen to the observed outcome under the same observed treatment regimen. We may succinctly write the sequential consistency assumption as $Y_j^g = Y_j^{\bar{a}}$ where $g = \bar{a}$. We use the potential framework to formulate the causal models for $Y_j^{\bar{a}}$ in which we estimate the diabetes

care provisions with respect to causal interventions $\bar{a}$. We assume that the censoring mechanism C_{ij} is completely at random in which the censoring process is independent of discretized time points T_{ij} and longitudinal outcome Y_{ij}, conditioned on observed cumulative history $\bar{\mathcal{H}}_{ij}$ as $C_{ij} \perp\!\!\!\perp \{Y_{ij}, T_{ij}\} | \bar{\mathcal{H}}_{i,j-1}$.

Model-based dynamic estimation

We use the directed acyclic graph (Figure (4.1)) to encode the relationships among time-dependent treatment process A_j, time-varying covariates L_{jk}, time-varying outcome-predictors M_{jk}, baseline covariates X_{jk} and repeated-measures outcome Y_j. We may use the directed acyclic graph to describe the treatment-confounder feedback, denoted using red edges in Figure (4.1) in which the past treatment A_{j-1} affects the current confounder L_{jk}, and the current confounder L_{jk} in turn affects the current treatment A_j. In traditional context, we account for treatment-confounder feedback using G-methods (e.g. marginal structural models or G-computation) (Naimi et al., 2017). In this article, we describe the treatment-confounder feedback using recurrent prescriptions (discretized annually) for glucose-lowering medications and appropriate time-dependent confounding features (e.g. 100 most common diagnostic ICD9 CPP codes and ATC codes).

Figure 4.1: Directed acyclic graph with time-dependent treatment-confounder feedback



Since we are interested in the causal estimation of the treatment process A_{ij} onto outcome process Y_{ij} in the presence of treatment-confounder feedback, we encode the marginal structural model with respect to time-varying covariates $L_{i,j-1,k}$ and Y_{ij-1} as

$$\Psi^{\bar{a}}_{ij} = Pr(Y^{\bar{a}}_{ij} | \bar{A}_{i,j-1}, \bar{M}_{i,j-1}) = \Phi\left(\bar{a}_{ij}, m_{i,j-1,k}\right) \tag{4.2}$$

where $\Phi(\cdot)$ denotes some arbitrary marginal function of outcome process with respect to time-dependent covariates $L_{i,j-1,k}$ and Y_{ij-1}. We note the exclusion of time-dependent confounders in Equation (4.2) because this may bias the direct or indirect treatment effects in the longitudinal causal structure (Robins et al., 2000). In similar fashion, we encode the treatment model with respect to time-dependent covariate process as

$$Pr(A_{ij}|\bar{\mathcal{H}}_{i,j-1}) = \Omega\left(x_i, l_{ijk}, y_{ij}, a_{i,j-1}\right) \tag{4.3}$$

where $\Omega(\cdot)$ denotes some arbitrary function of treatment process. We employed the cumulative-time weight functions to marginalize the outcome process with respect to the time-varying covariates process. Both equations (4.2) and (4.3) describe arbitrary functions in which the statistical learning algorithms (e.g. ensemble-trees or neural networks) allow for interaction, non-linear and higher order effects to approximate the intricate functions (Rose, 2013).

Dynamic estimands using causal treatment modalities

We evaluate the hypothetical treatment contrast using "*pairwise estimands*" as a change in probability (i.e. causal risk difference) of receiving optimal diabetes provision within a calendar year with respect to two mutually exclusive treatment modalities. We may formalize the pairwise estimands for hypothetical treatment modality $\bar{a}$ under the dynamic treatment regimen as

$$\text{Average treatment effect} = \left(\Psi_{ij}^{\bar{a}} - \Psi_{ij}^{\bar{a}'}\right) \times 100\% \tag{4.4}$$

where $\Psi_{ij}^{\bar{a}}$ characterizes the hypothetical outcome probability with respect to treatment modality $\bar{a}$, and where $\bar{a} \neq \bar{a}'$. We formulate the hypothetical treatment modality using multinomial propensity score equations with $2^3 = 8$ possible treatment combinations within each calendar year. Since the hypothetical treatment modalities are indexed with respect to longitudinal calendar year (i.e. $j = \{2016, 2017, 2018\}$), this may give rise to $(2^3)^3 = 512$ possible treatment regimen. We restrict the hypothetical pairwise estimands to homogeneous treatment modalities with respect to longitudinal follow-up (e.g. only Metformin in 2016, 2017, 2018). This simplification of counterfactual treatment modalities lead to the comparison of $\binom{8}{2} = 28$ pairwise estimands, and thereby mitigating the combinatorial explosion of hypothetical treatment regimen indexed with respect to calendar year (i.e. $(2^3)^3 = 512$ possible treatment regimen).

Stabilizing weight function

We introduce the stabilizing weight function to eliminate the associations between the time-varying covariate process L_{ijk} and time-varying outcome process Y_{ij}. Regardless of the functional relationships imposed using the statistical learning algorithms, we may describe the stabilizing weight function with respect to longitudinal treatment process A_{ij} as

$$SW_{ij}^{\bar{A}} = \prod_{t=1}^{j} \frac{Pr(A_{it}|\bar{\mathcal{H}}_{i,t-1}/\{L_{i,j-1,k}, Y_{ij-1}\})}{Pr(A_{it}|\bar{\mathcal{H}}_{i,t-1})} \tag{4.5}$$

where the numerator $Pr(A_{ij}|\bar{\mathcal{H}}_{i,j-1}/\{L_{i,j-1,k}, Y_{ij-1}\})$ describes the stabilizing factor with the exclusion of time-dependent covariates while the denominator $Pr(A_{ij}|\bar{\mathcal{H}}_{i,j-1}) \equiv Pr(A_{ij}|L_{i,j-1}, Y_{ij-1})$ describes

the inverse probability of treatment assignment with the inclusion of time-dependent covariates. Pajouheshnia et al. (2020) used an inverse probability censoring weights to account for informative censoring in estimating the treatment-naïve risk. The application of the censoring weights was not necessary since the censoring mechanism was assumed to be completely at random with respect to the discretized time points and longitudinal outcome, conditioned on appropriate covariate history. Only the stabilized inverse probability treatment weights (with the calibrated restrictions) are used to create the pseudo-population in which the time-dependent treatment process A_{ij} becomes unconfounded. Similar to Dong (2021), we truncate the stabilizing weight function and the calibrated weight function at 0.5% and 99.5% quantiles to improve the estimation of the marginal treatment effects (Xiao et al., 2013).

Calibration of stabilizing weight function

In survey sampling, the calibration of weight functions are performed to integrate the auxiliary information in which the distance between the initial weights and final weights is minimized subject to calibrating restrictions (Deville and Särndal, 1992). We introduce the calibration framework in this article to improve the finite-sample covariate balance of the stabilizing weight function (Yiu and Su, 2020). In particular, we formulate the calibration procedure for the stabilizing weight function to improve the covariate balance with respect to the observed time-dependent covariates $L_{ik,t-1}$ as

$$\sum_{i=1}^{n}\sum_{j=2016}^{2019} SW_{ij}^{\bar{A}}(\lambda)\sum_{t=1}^{j}\left[(A_{it}-\hat{e}_{it}^{A})\times L_{ik,t-1}\right]=0 \tag{4.6}$$

where $SW_{ij}^{\bar{A}}(\lambda)=SW_{ij}^{\bar{A}}\times exp(K\lambda)$ denotes the calibrated stabilized weights with the unknown parameter λ and data-dependent covariate restrictions in matrix K. In equation (4.6), we notice that the residual of propensity scores (i.e. $(A_{it}-\hat{e}_{it}^{A})$, where $\hat{e}_{it}^{A}=Pr(A_{ij}|\bar{\mathcal{H}}_{i,j-1})$) must be orthogonal to $L_{ik,t-1}$ since $SW_{ij}^{\bar{A}}(\lambda)$ are constrained to be non-negative. This orthogonality constraint ensures that the propensity score residuals are linearly independent with respect to the time-varying covariates $L_{ik,t-1}$ in high-dimensional Euclidean space (Rodgers et al., 1984).

Although the stabilized weights in the pseudo-likelihood function of marginal structural models satisfy the property of unity mean (i.e. $E(SW_{ij}^{\bar{A}})=1$ at each time-point j) (Hernán and Robins, 2006), this property is not guaranteed to hold for calibrated stabilized weights (Yiu and Su, 2018b). In additional to the time-dependent covariate balancing constraints (above in equation (4.6)), we also impose the restriction for average calibrated weights to be equal to one at each time-point j as

$$E(SW_{ij}^{\bar{A}}(\lambda))=\frac{1}{n}\sum_{i=1}^{n}SW_{ij}^{\bar{A}}(\lambda)=1. \tag{4.7}$$

We used the calibrated weights satisfying equation (4.6) and (4.7) to construct the pseudo-population for the longitudinal diabetes cohort and to assess the covariate balance in hypothetical treatment regimes with respect to metformin, sulfonylurea and SGLT-2i The constrained optimization is implemented using the Barzilai-Borwein gradient method in R software (Varadhan and Gilbert, 2009).

4.4.3 Machine learning algorithms

In similar spirit to Blakely et al. (2020) and Karim et al. (2017b), our aim is to estimate the marginal means using the machine learning algorithms. We are interested in conducting supervised machine learning using a collection of mainstream statistical learning algorithms including least absolute shrinkage and selection operator, ridge regression, elastic net, random forest, gradient boosting machine and neural network. We provide a brief summary of each base learner in the Supplementary Section (4.10). We used this collection of machine learning algorithms to build a stacked hypothetical estimation using the SuperLearner.

SuperLearner

The SuperLearner algorithm combines the estimation from individual base learner to create a stacked estimation (Breiman, 1996). Since both causal effects and longitudinal estimation (in the context of machine learning) can be described as an estimation problem, the idea is to further improve the causal estimation using the SuperLearner in which the stacked estimand is indexed with respect to multiple base learners (Van der Laan and Rose, 2011). In many instances, the SuperLearner algorithm outperforms individual base learners (e.g. regularization methods, ensemble-based trees or deep learning using neural networks) to generate the most optimal system for estimation (Van der Laan et al., 2007). Unlike the earlier ensemble based methods (e.g. tree-based), the stacked ensembles in SuperLearner algorithm represents a *"diverse group of strong base learners"* with parametric, semi-parametric or non-parametric assumptions (Boehmke and Greenwell, 2019). In similar spirit to Rose (2013), we formulate the SuperLearner algorithm in the context of hypothetical estimation using the following steps:

1. We select the brute-force configuration of the entire hyperparameter grid search for a collection of machine learning algorithms: (i) lasso regression, (ii) ridge regression, (iii) elastic net regression, (iv) random forest, (v) gradient boosting machine, (vi) neural network.

2. We apply the patient-level data split on training sample to create 10 mutually exclusive and exhaustive blocks of equal (or approximately equal) size. We apply the clustered 10-fold CV in which the cumulative-time product treatment weights were preserved for each patient within 10 blocks.

3. We fit each machine learning algorithm (i)-(vii) using 10-fold CV with calibrated weights. We use the validation set in the training sample (using 10-fold CV) to predict the probability of diabetes provision $\Psi^{\bar{a}}_{ij}(W)$ for i^{th} individual at j^{th} time-point for w^{th} machine learning algorithm.

4. We gather the estimated probabilities $\Psi^{\bar{a}}_{ij}(W)$ for the entire training set and then estimate the CV MSE for each machine learning algorithm w (see equation (4.10)).

5. We estimate the optimal weight combinations for machine learning algorithms indexed with respect to the weight vector α using the non-negative least square estimation as

$$\Psi^{\bar{a}}_{ij}(SL) = \sum_{l=1}^{L} \alpha_l \Psi^{\bar{a}}_{ij}(W)$$

where α_l characterizes the SuperLearner weights and $\Psi^{\bar{a}}_{ij}(SL)$ denotes the predicted probability of the SuperLearner.

6. We use the estimated weights for each machine learning algorithm in the SuperLearner to generate estimation in the held-out test sample.

Since the estimation problem of diabetes provision (in next calendar year) can be considered as repeated-measures problem, we perform sample-split on each independent patient units (Balzer and Petersen, 2021). Splitting the training and test set at the patient level (rather than at the repeat observations) allows us to preserve the cumulative-time products of stabilized weight function within each sample split, and eliminate the time-dependent confounding process. We estimate the counterfactual probabilities (in the test sample) with respect to eight treatment modalities (separately) for each base learner with non-negative weight contributions to the SuperLearner. The counterfactual probabilities of the base learners are then amalgamated using the non-negative least squares to generate stacked estimations for each counterfactual treatment. We may describe the variance of the causal risk difference as

$$\begin{aligned} Var(\Psi_{ij}^{\bar{a}} - \Psi_{ij}^{\bar{a}'}) &= Var(\Psi_{ij}^{\bar{a}}) + Var(\Psi_{ij}^{\bar{a}'}) - 2Cov(\Psi_{ij}^{\bar{a}}, \Psi_{ij}^{\bar{a}'}) \\ &= Var(\Psi_{ij}^{\bar{a}}) + Var(\Psi_{ij}^{\bar{a}'}) \end{aligned}$$

where $\Psi_{ij}^{\bar{a}}(sl) \perp\!\!\!\perp \Psi_{ij}^{\bar{a}'}(sl)$. The bootstrap samples (with replacement) for 100 iterations are used to generate the 95% confidence intervals of the average treatment effect with respect to hypothetical treatment modalities (Wasserman, 2013).

4.4.4 Implementation of Machine Learning pipelines

We describe the machine learning pipelines using the generation of longitudinal diabetes cohort and its data splitting into training and test sample, followed with the discussion on the marginalization of covariate process to generate hypothetical estimation. We describe the criteria for tuning the hyperparameter grid search of machine learning algorithms, and criteria to assess the performance of machine learning algorithms using the appropriate evaluation metrics.

Generation of longitudinal diabetes cohort

We construct a longitudinal diabetes cohort in which patients are enrolled when the following conditions were satisfied: (i) patients are at least 40 years of age as of January 1^{st} of each index year; (ii) patient has an indication in EHRs corresponding to diabetes; (iii) research quality criteria for EHRs is satisfied (Tu et al., 2020a); (iv) patient must have at least one visit recorded in billing or encounter fields within calendar year; (v) exclusion of Type I diabetes patients (Weisman et al., 2020). The age restriction for condition (i) is in agreement with the diabetes provision guidelines (Rigobon et al., 2019), while condition (ii) is borrowed from earlier work on diabetes phenotype (Williamson et al., 2014). We impose condition (iv) as an interval censoring mechanism to account for the visit process (i.e. no EHR data are collected in the absence of visit within a calendar year). In addition to interval censoring (Zhu et al., 2017), we impose administrative censoring mechanism where the patients are censored at the end of the study period (December 31, 2019). We use this open cohort design with time-dependent risk-set to make hypothetical estimation of diabetes care provision. We enrich the prediction matrix with elements captured from electronic health records including (i) patient demographics, (ii) diabetes

medication classes, (iii) lab characteristics, (iv) vaccination, (v) lifestyle information, (vi) ICD-9 billing codes, (vii) ICD-9 CPP codes, (viii) ATC codes and (ix) OHIP codes.

Data splitting

During the data pre-processing step, it is necessary to prevent *"data leakage"* whereby the information may propagate outside the training set (Kaufman et al., 2012). A trivial example of data leakage may include the use of individual diabetes care elements (e.g. blood pressure count) of target output (i.e. composite binary index of "diabetes provision") as inputs. We mitigate the possibility of *"data leakage"* with two data pre-processing steps. First, we generate a dynamic cohort in which the predictors (including the individual elements of diabetes care) are time-lagged with one calendar year with respect to the composite binary outcome of "diabetes provision". Second, we perform the data splitting step for training sample and testing sample prior to re-sampling iterations of machine learning algorithms. The second step ensures that we did not screen for any strong predictors prior to 10-fold cross-validation (Friedman et al., 2001). Using the total number of unique patients as the sampling unit, we spit the longitudinal diabetes cohort data as 80% training sample and 20% test sample. The 80% training sample is further split into 10-fold cross-validation sample to generate the appropriate diagnostic metrics for machine learning algorithms (as described further in Section (4.4.4)).

Feature Engineering

We capture several elements of primary care electronic health records, and incorporate them as high-dimensional prediction matrix using *"one-hot"* (dummy) encoding. In particular, we implement the feature engineering as boolean design matrix for the following elements in electronic health records using the annual calendar-time discretization: (i) demographics (X_{ik}): age group (as of January 1 of index year), sex, income quintiles, rurality, deprivation index, ethnic concentration; (ii) laboratory requisition (M_{ijk}): hemoglobin test, hemoglobin A1c test, low and high density lipoprotein test, serum cholesterol test; thyroid-stimulating hormone test, fasting blood glucose test, prostate antigen test, human chorionic gonadotropin (HCG) test, international normalization ratio (INR) test, 25-Hydroxy Vitamin D test, Hepatitis B Blood test; (iii) vaccination and lifestyle (M_{ijk}): influenza vaccination, alcohol consumption, smoking status; (iv) diabetes medications (A_{ij}): Metformin, Sulfonylurea, Sodium-Glucose Co-transporter-2 (SGLT-2i inhibitors; (v) 100 most common diagnostic International Classification of Diseases v9 (ICD-9) billing codes (M_{ijk}); (vi) 100 most common diagnostic ICD-9 cumulative patient profile (CPP) codes (L_{ijk}); (vii) 100 most common medications using Anatomical Therapeutic Chemical Classification (ATC) nomeclature (L_{ijk}); (viii) 100 most common Ontario Health Insurance plan (OHIP) billing codes (M_{ijk}). The feature engineering of these predictors is implemented using binary encoding scheme and it may be described as

$$\text{Feature}(t) = \begin{cases} 1 & \text{if present within calendar year } t \\ 0 & \text{if absent within calendar year } t \end{cases} \tag{4.8}$$

where we index each feature with respect to discrete calendar year t. We construct a rank-ordered (time-invariant) index for "100 most common" features using the overall frequency count in NDR. The rank-ordered ICD-9 diagnostic codes, ATC codes and OHIP billing codes remains unchanged with respect to each index year from 2016 to 2019.

Marginalization of the covariate process

We apply the machine learning algorithms using two models: (i) treatment model to estimate the probability of receiving post-baseline treatment; (ii) an outcome model for "diabetes care provision" in next calendar year using the inverse probability treatment weights. Prior to the outcome model, we eliminate the associations between covariate process L_{ijk} and treatment process A_{ij} using the cumulative-time product weight function with calibrated restrictions as described in Section (4.4.2). The marginalization with respect to covariate process (or equivalently the elimination of the time-varying confounding process) allow us to generate the hypothetical estimation for diabetes care provision. Lee et al. (2010) used machine learning methods to estimate the propensity scores for binary treatment assignments, and showed a reduction in bias and mean square error (MSE) for causal estimands using machine learning methods as compared to simple logistic regression model. As an extension, McCaffrey et al. (2013a) estimated propensity score for multiple treatment assignment using the generalized boosted models. Building on McCaffrey et al. (2013a), we applied the ensemble-based gradient boosting trees to compute the propensity scores for multinomial prescriptions of glucose-lowering medications: metformin, sulfonylurea and SGLT-2i, and their corresponding combinations. Using the estimated propensity scores, we build the stabilized weight functions as discrete cumulative-time product to account for the time-dependent confounding and then enforce the calibrated constraints to improve covariate balance in the pseudo-population of longitudinal diabetes cohort (as described in Section (4.4.2)).

Tuning hyperparameter grid search

We construct a hyperparameter grid for each machine learning algorithm using the factorial configuration (as listed in Table 4.1). We apply the hyperparameter grid of gradient boosting machine on the treatment process (i.e. glucose lowering medications) to compute the cumulative-time product weights (described in Section 4.4.2). In machine learning applications, there exist multiple criteria to tune the hyperparameters including MSE, one-standard error and area under the curve (Friedman et al., 2001). The one-standard error criteria may be used to select a parsimonious model in relation to more complex model often selected using the minimization of the MSE. We apply the criteria for the minimization of MSE to achieve improved estimation of multinomial propensities of glucose-lowering treatment assignment, which are then transformed into cumulative-time product weight functions.

Once the calibrated weights are estimated, we brute-force the entire hyperparameter grid of each statistical learning algorithm for the stacked estimation using the SuperLearner. In particular, we apply the entire hyperparameter grid of each base learner to the training (and held-out 10-fold cross-validation) set using the cumulative-time product weights. We then stack the cross-validated prediction in the training set and externally validate the performance of the SuperLearner (see Section (4.4.3)).

Standardized mean difference

We apply the minimization of the MSE in the test sample as the evaluation metric to select the most optimal hyperparameter configuration for the treatment model using the gradient boosting machines. Once the optimal configuration of hyperparameters is selected for the treatment model, we evaluate the covariate balance in the pseudo-population based on standardized mean difference (SMD). The covariate

balance is assessed for k time-dependent covariates used in the treatment model and can be defined as

$$\text{SMD}_{jk} = \frac{\hat{p}_t - \hat{p}_c}{\sqrt{\frac{(\hat{p}_t)(1-\hat{p}_t)+(\hat{p}_c)(1-\hat{p}_c)}{2}}} \tag{4.9}$$

where $\hat{p}_t$ denotes the weighted average of treatment drop-in cohorts while $\hat{p}_c$ denotes the weighted average for treatment naïve cohort. The denominator in equation (5.28) correspond to pooled standard deviation of treatment and control regimen. The covariate balance in the pseudo-population is assessed using the difference in prevalence measured relative to the units of the pooled standard deviation (Austin, 2009).

Mean square error

We use the mean square error (MSE) to assess the performance of each base-learner with non-negative weight contribution to the SuperLearner prediction. We use the predicted probabilities $\Psi_{ij}(W)$ to estimate the CV MSE for each machine learning algorithm w as

$$\text{CV } MSE(w) = \frac{\sum_{i=1}^{n} \sum_{j=2016}^{2019} \mathbb{1}(\text{Visit count} \geq 1)_{ij} \times (Y_{ij} - \Psi_{ij}(W))^2}{N} \tag{4.10}$$

where $\mathbb{1}(\text{Visit count} \geq 1)_{ij}$ describes the interval censoring with respect to the visit process within each calendar year t, Y_{ij} denotes the diabetes provision for individual i at j^{th} time-point, and N denotes the sample size of training set.

Table 4.1: Hyper-parameter grid of machine learning algorithms for causal prediction of diabetes provision

Type	**Machine learning algorithm (n)**	**Hyperparameter grid search**
Regularization methods	Lasso regression (8)	$\alpha = \{1\}$ $log(\lambda) = \{-12 \text{ to } -5 \text{ by } 1\}$
	Ridge regression (8)	$\alpha = \{0\}$ $log(\lambda) = \{-12 \text{ to } -5 \text{ by } 1\}$
	Elastic net regression (16)	Elastic net mixing parameter $\alpha = \{0.2 \text{ to } 0.8 \text{ by } 0.2\}$ $log(\lambda) = \{-12 \text{ to } -6 \text{ by } 2\}$
Ensemble-tree methods	Random forest (12)	# of trees = {100, 200} # of sampled predictors= {15, 30} Maximum depth={4, 8, 16} Sample with replacement={True}
	Gradient boosting machine (12)	# of trees = {100, 200} Maximum tree depth={4, 8, 16} Learning parameter $\eta = \{0.01, 0.001\}$
Non-parametric methods	Neural network (8)	# of units in the hidden layer={128, 256} Weight decay = {0, 0.05, 0.10, 0.15} Activation function= {Logistic}
Stacking	Superlearner (1)	Meta-algorithm={non-negative least square (NNLS) using logistic regression}

(n) Total number of hyperparameter grid combinations.

4.5 Results

We describe the results in three subsections: (i) longitudinal cohort description using annualized aggregation; (ii) covariate balance using cumulative product time weights; (iii) hypothetical predictions for diabetes provision using the SuperLearner.

4.5.1 Cohort description

We describe the distribution of diabetes provision in 2017, 2018 and 2019 using the patient demographics (age group, sex), geographical characteristics (income quintiles, rurality), and treatment modalities (Metformin, Sulfonylurea and SGLT-2i in Table (4.2), (4.3) and (4.4), respectively. We noticed an improvement in diabetes provision with respect to increase in age groups (with the exception for 80+ years). Male patients tend to receive improved diabetes care with higher prevalence than female patients for each calendar year. A slight increase in prevalence of diabetes provision was detected in lowest income quintiles while a non-distinguishable difference in prevalence of diabetes provision was captured with respect to urban or rural regions.

The adequate prevalence of diabetes provision was consistently lower (for three consecutive years) among patients who did not receive a prescription for Metformin, Sulfonylurea and SGLT-2i Any combination of prescriptions related to glucose-lowering medications led to improved prevalence of adequate diabetes provision in next calendar year. Patients who received diabetes screening services in previous year were likely to receive better diabetes provision in next calendar year: (i) two or more primary care visits (77% vs 56%), (ii) two or more blood pressure count (84% vs 61%), (iii) two or more weights recorded (87% vs 67%), (iv) two or more HbA1c test (87% vs 60%), (v) one or more lipid panel test (82% vs 64%), (vi) one or more ACR test (87% vs 69%), (vii) one or more eGFR test (81% vs 57%), (viii) one or more statin prescription (85% vs 65%).

4.5.2 Covariate balance

The stabilized weight function is used to construct a pseudo-population in which the balance is achieved with respect to the distributions of the time-dependent covariates in each treatment regimen. Figure (4.2) describes the scatterplot between stabilized weights and calibrated weights for eight treatment modalities. The pseudo-population in terms of covariate balance was further improved using the calibrated stabilized weights. The constrained optimization was applied on stabilized product weights to estimate the calibrated weights. The side panels in Figure (4.2) show the density plots of stabilized and calibrated weights with respect to each treatment modality. The interquartile range of (cumulative-time) stabilized weights ranged was 0.111 and 0.395 with mean value 0.270 while the interquartile range of calibrated weights was 0.308 and 1.363 with mean value 0.890. The correlation between the stabilized weights and calibrated stabilized weights was noted to be 0.725 (95% CI: 0.722- 0.727).

SMD is used to describe the covariate balance in each treatment drop-in cohort (using a combination of Metformin, Sulfonylurea and SGLT-2i with respect to the treatment naïve cohort (i.e. no treatment regimen). Figure (4.7) describes the covariate balance for each time-dependent covariate L_{ijk} (i.e. ICD9 CPP codes and ATC medication codes) using the calibrated weights. Most of the covariates were within the ± 0.20 caliper range with few notable exceptions. Out of 197 time-dependent covariates, the calibrated weights contained 182 covariates (92.4%) within ± 0.20 caliper range for seven treatment drop-in cohorts in relation to treatment naïve cohort.

Table 4.2: Diabetes provision in 2017 using 2016 predictors in Diabetes Action Canada Repository

			Diabetes provision				
Characteristics			**Less than adequate care**		**Adequate to optimal care**		**Total**
Age groups			**N**	**Row Percent (%)**	**N**	**Row Percent (%)**	**N**
40-49 years			1,624	38.8%	2,566	61.2%	4,190
50-59 years			2,863	29.3%	6,898	70.7%	9,761
60-69 years			2,836	22.0%	10,046	78.0%	12,882
70-79 years			1,893	17.4%	8,978	82.6%	10,871
80+ years			1,479	24.5%	4,548	75.5%	6,027
Sex							
Female			5,567	26.3%	15,605	73.7%	21,172
Male			5,128	22.7%	17,431	77.3%	22,559
Income Quintiles							
1			2,390	23.2%	7,908	76.8%	10,298
2			2,088	24.5%	6,425	75.5%	8,513
3			1,883	24.8%	5,702	75.2%	7,585
4			1,682	24.0%	5,322	76.0%	7,004
5			2,135	25.8%	6,134	74.2%	8,269
Missing			517	25.1%	1,545	74.9%	2,062
Rurality							
0			9,163	24.4%	28,349	75.6%	37,512
1			1,532	24.6%	4,687	75.4%	6,219
Metformin	**Sulfonylurea**	**SGLT-2i**					
0	0	0	7,521	34.0%	14,607	66.0%	22,128
0	0	1	57	15.0%	324	85.0%	381
0	1	0	225	17.4%	1,068	82.6%	1,293
0	1	1	29	20.9%	110	79.1%	139
1	0	0	2,075	15.4%	11,385	84.6%	13,460
1	0	1	140	13.4%	904	86.6%	1,044
1	1	0	578	12.6%	4,015	87.4%	4,593
1	1	1	70	10.1%	623	89.9%	693
Total			10,695	24.5%	33,036	75.5%	43,731

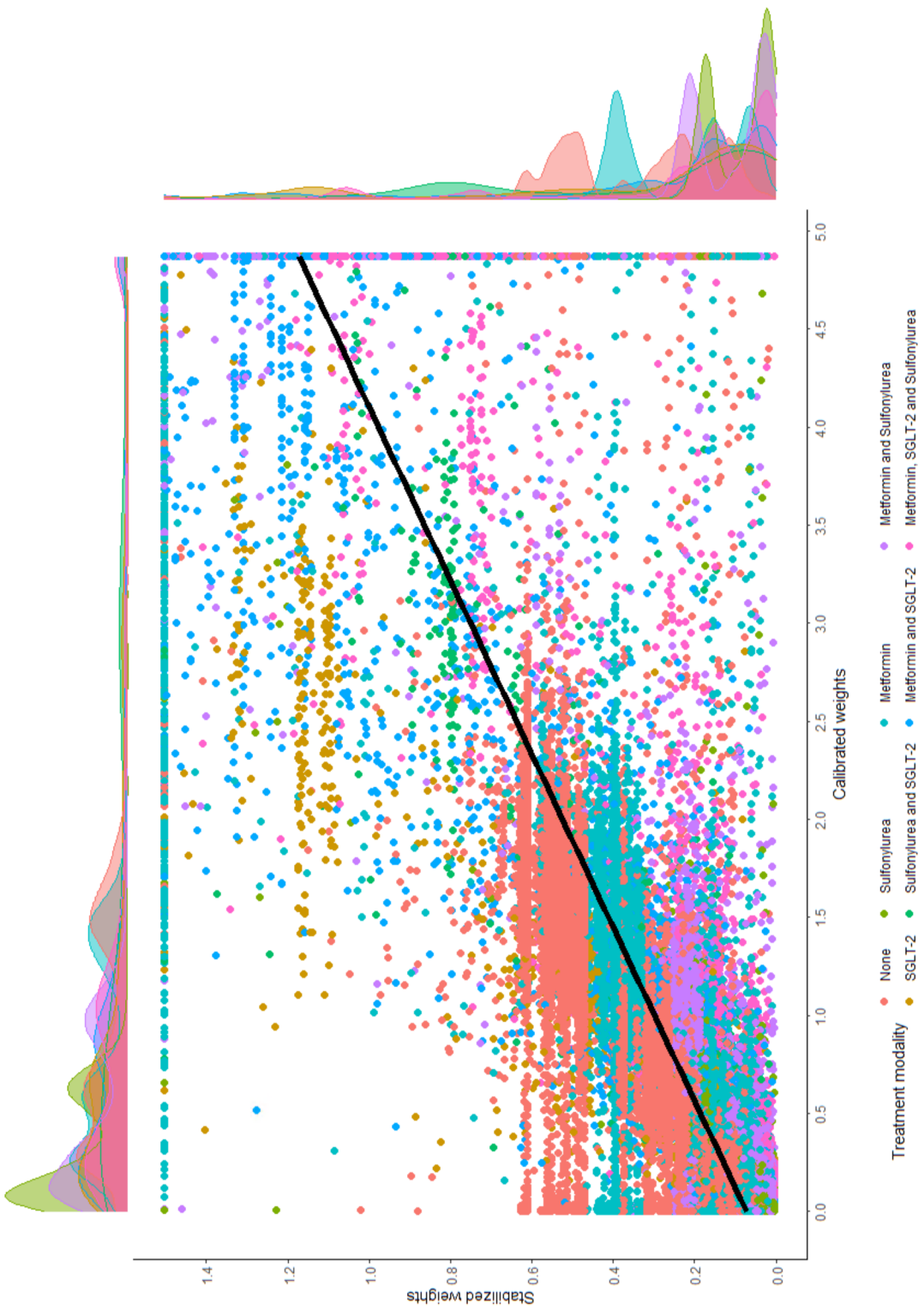


Figure 4.2: Scatterplot of stabilized and calibrated treatment weights

4.5.3 Brute-force hyperparameters grid search for base-learners

Figure (4.3) describes the predicted probability of adequate diabetes care for each base learner with respect to the binary outcome of diabetes care provisions. For example, the predicted probability close to the extreme end of one correspond to higher chance of receiving adequate diabetes care provisions as predicted by the base learner.

4.5.4 Stacked estimation using the SuperLearner algorithm

Figure (4.4) shows the magnitude of non-negative least squares coefficients to generate the stacked estimation of the SuperLearner. The two panels in Figure (4.5) describe the predicted probabilities of adequate diabetes care provisions based on the SuperLearner algorithm in 10-fold cross-validated training sample and test sample. The SuperLearner had area under the receiver operating curve (AUROC) estimate of 0.761 (95% 0.758 - 0.765) in the training sample and 0.773 (95% 0.766 - 0.780) in the test sample. The improved AUROC estimates of SuperLearner algorithm in relation to the base learners demonstrated how the amalgamation of statistical learning algorithms using the non-negative least square estimation may improve the diagnostic properties of causal estimation.

4.5.5 Causal estimation

We generated the causal estimation with respect to homogeneous treatment modalities in 2017, 2018 and 2019. Figure (4.6) describes the average treatment effect using causal risk difference between two mutually exclusive treatment modalities in the test sample. In general, any combination of glucose lowering medications (i.e. metformin, sulfonylurea, or SGLT-2i) led to improved diabetes care provisions in relation to treatment naïve modalities. As an example, the hypothetical treatment modality of metformin in each calendar year (i.e. 2016, 2017 and 2018) improved diabetes care provisions by 1.6% (95% CI: 1.0% - 2.3%) in relation to the hypothetical treatment naïve cohort.

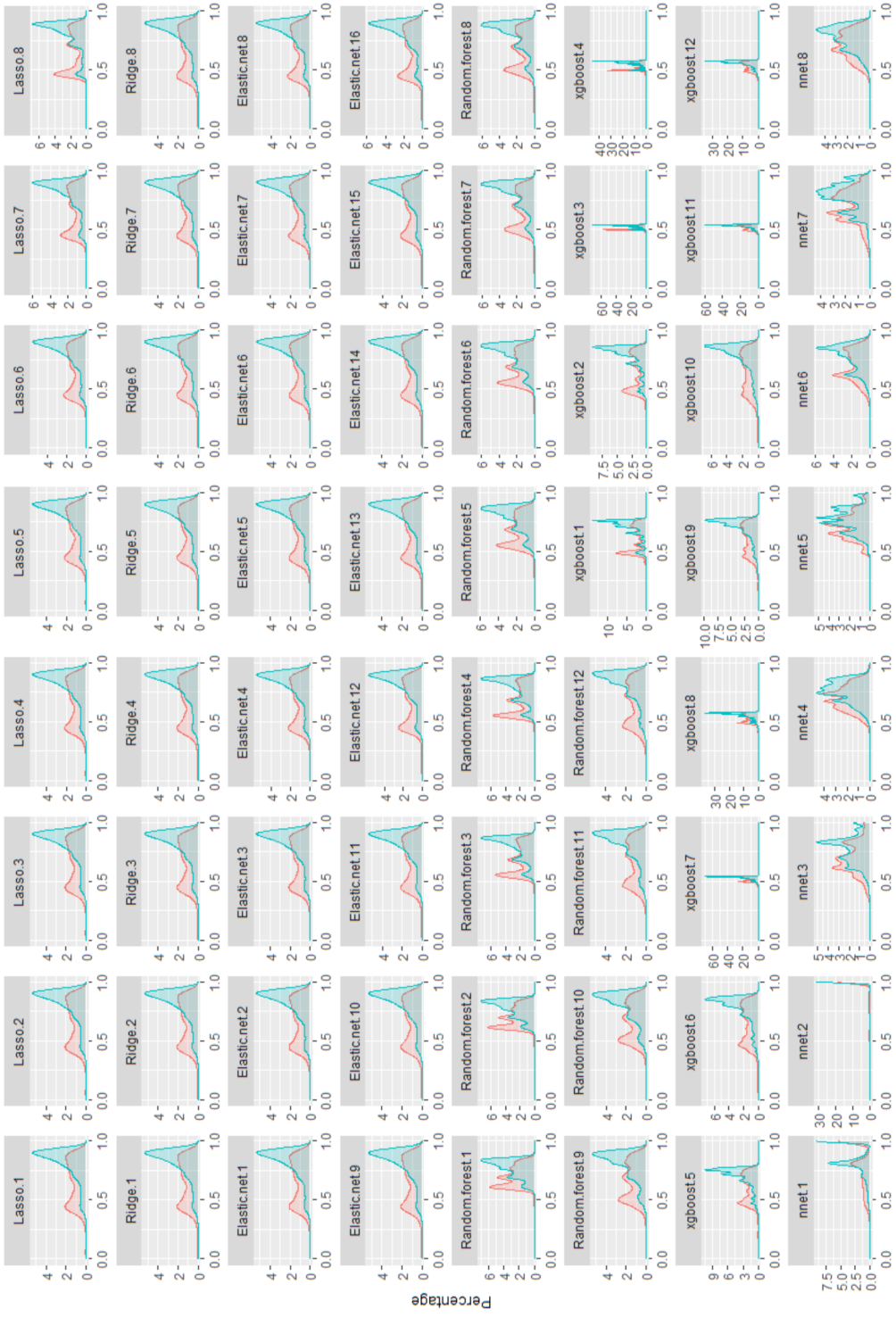


Figure 4.3: Ten-fold cross-validation predicted probabilities of diabetes care provisions for each base learner

Figure 4.4: Non-negative coefficients of base learners in 10-fold cross-validated training sample

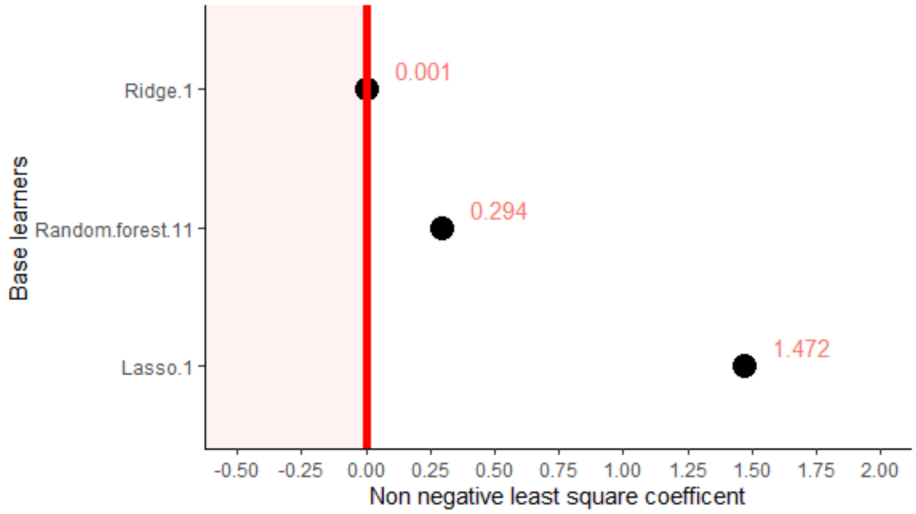


Figure 4.5: Predicted probabilities of adequate diabetes provisions for the SuperLearner in training sample and test sample

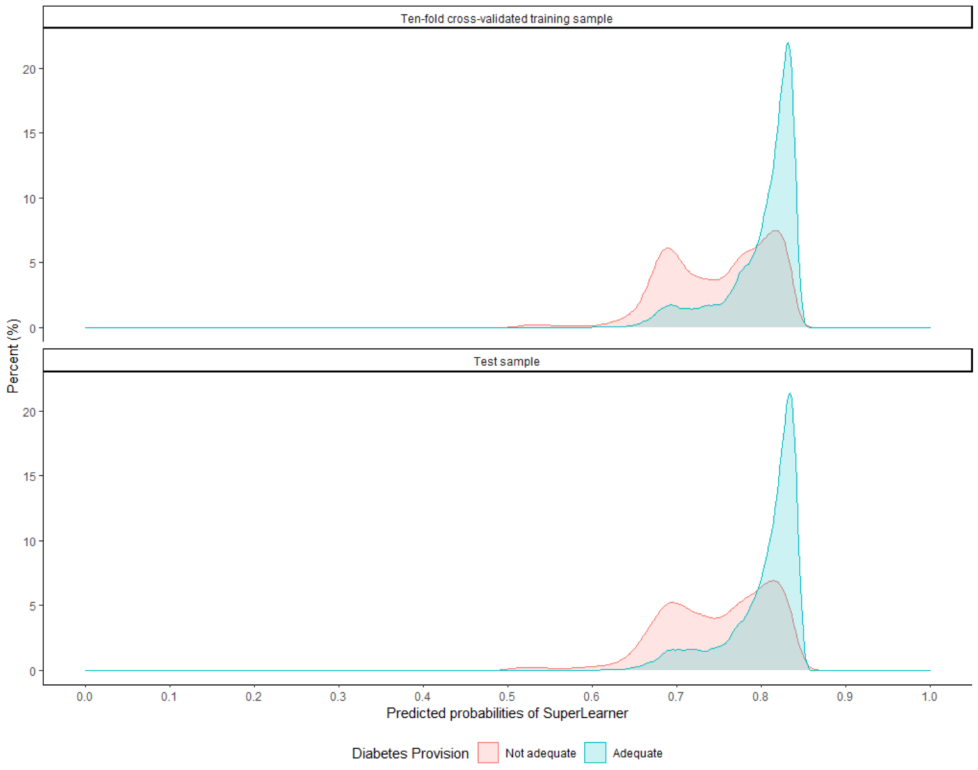
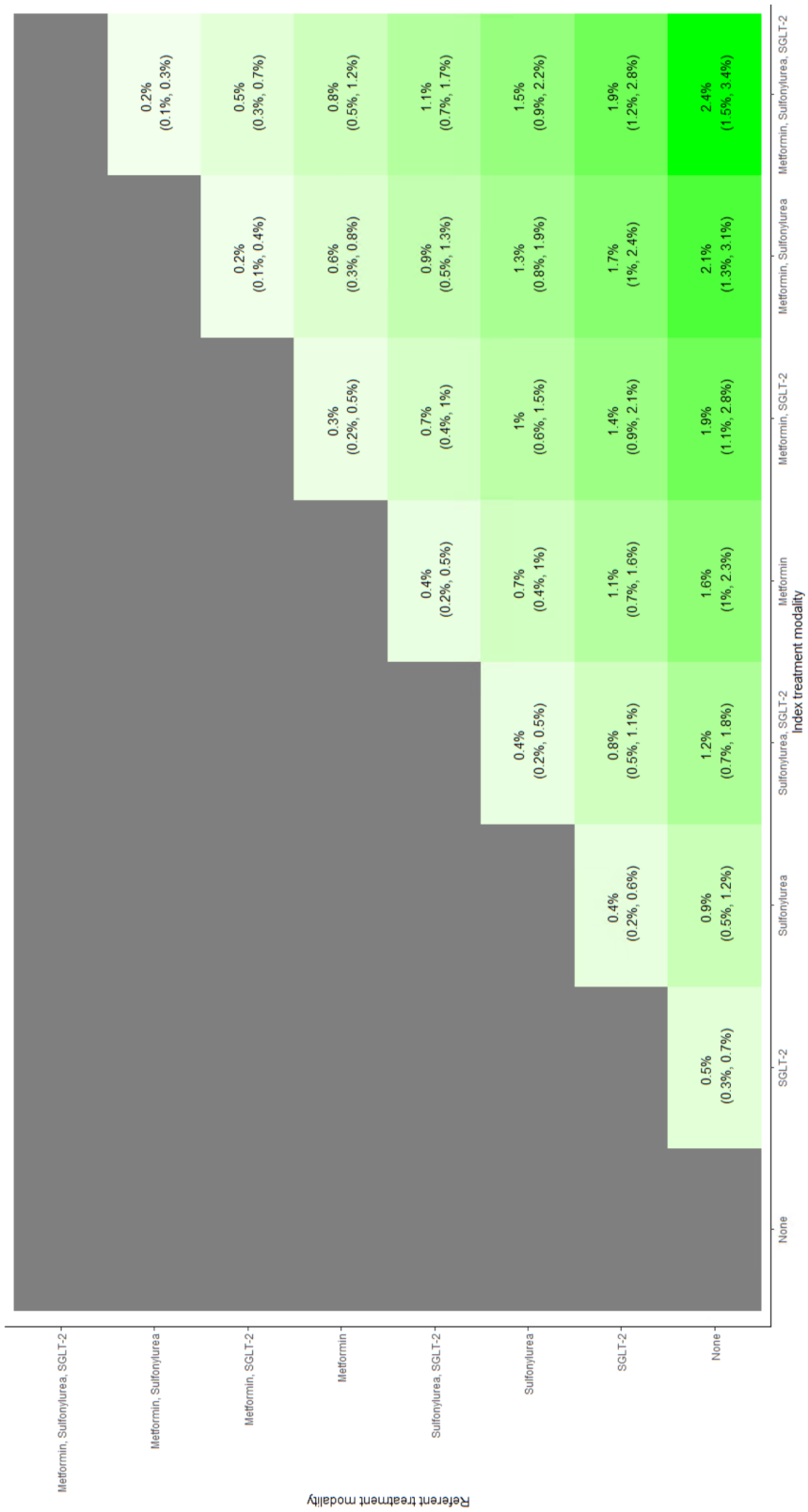


Figure 4.6: Hypothetical risk difference with 95% bootstrap confidence interval using the SuperLearner prediction in test sample



4.6 Discussion

There is a rich history for the application of statistical learning algorithms in the context of clinical epidemiology research of diabetes (Basu et al., 2020). However, less emphasis is placed on research using causal estimation in diabetes context using EHRs. The overarching aim of this article was to demonstrate how the causal estimation of diabetes care provisions (indexed with respect to glucose-lowering medications) can be applied using an ensemble of machine learning algorithms. Reasonable covariate balance was achieved using the calibrated weights with respect to time-dependent covariate distributions in eight treatment modalities. Our results indicated that hypothetical treatment regimens (with respect to metformin, sulfonylurea and SGLT-2i) may improve diabetes care provisions in next calendar year while accounting for time-dependent covariates using the calibrated weights. These findings may help to inform the clinical practice guidelines for diabetes patients in which the allocation of primary care services may be designed proactively (Ivers et al., 2019). For example, if we may hypothetically predict which patients with type 2 diabetes, under normal circumstances, would be less likely to attend for care, do their laboratory tests and/or be prescribed recommended medications, we may better plan outreach programs using virtual care in this pandemic (Kiran et al., 2020).

Kohane et al. (2021) describe six aspects of critically appraising EHR research studies: (i) data completeness, (ii) data collection and handling (e.g. harmonization), (iii) data type, (iv) robustness of methods against EHR variability, (v) transparency of data and analytic code, (vi) multidisciplinary collaborations. We incorporated these elements in this study with the hope that it will foster rigor, quality and reliability for future studies using primary care EHRs. In similar spirit to Kohane et al. (2021), we describe the completeness of EHR features (e.g. specific lab test, OHIP billing codes, diagnostic ICD-9 codes) with regards to the absence or presence of specific feature within a discrete calendar year. Unlike other EHR studies, this study only considered structured EHR information with minimal risk of patient identifiers in relation to EHR studies using unstructured information (e.g. free-text for natural language processing task). During the data collection and harmonisation process, the de-identification procedures (with detailed documentation) are the cornerstone of building a national primary care chronic disease surveillance (e.g. diabetes) network in Canada (Keshavjee et al., 2011). We also strive for transparent data collection, and data harmonization procedures at NDR, with appropriate details provided on `https://repository.diabetesaction.ca/`. We limited the scope of this study to EHRs within Ontario (using UTOPIAN and EON data at NDR) to ensure *"robustness of methods against EHR variability"*, as data extraction practices across multiple provinces in Canada are likely to impact the hypothetical estimation of machine learning algorithms due to the presence of data heterogeneity. This study involved *"multidisplinary collaborations"* (across clinicians, data managers, data scientists and statisticians) to gather the most clinically relevant information on EHR features and to construct a dynamic causal estimation tool (using the SuperLearner) for diabetes provision in the longitudinal cohort. The NDR is built with interdisciplinary expertise of health policy scholars, clinicians, data managers, data scientist and statisticians as well as routine engagement from community members, policy makers and stakeholders.

It is necessary to ground the application of statistical learning algorithms with the formal framework of counterfactuals in causal inference, as the methodological aspects of "*causal prediction models*" are further developed in the literature (Lin et al., 2021). Balzer and Petersen (2021) provide practical recommendations on how to integrate statistical learning algorithms with causal analyses, and we incorporated the recommended "*Causal Roadmap*" in this article. For example, it is necessary to state the research question with appropriate description of the target population, treatment modalities and primary out-

come. We encapsulated the longitudinal causal relationships, along with potential source of biases (e.g. time-dependent treatment-confounder feedback), in the directed acyclic graph (as shown in Figure 4.1). We acknowledged the necessity of the identifiability assumptions to warrant the interpretation of the causal estimands. Since time-dependent confounders exist as a mediating factor in recurrent treatment process and outcome process, we cannot adjust for the time-dependent confounders in the outcome model, and instead we must use the inverse probability treatment weights in marginal structural models (Xiao et al., 2010). We heed the advice of "*causal model neglect*" by carefully specifying the target parameter using the apriori clinical knowledge encoded in the directed acyclic graph before proceeding with the causal estimation using the statistical learning algorithms (Balzer and Petersen, 2021). To the best of our knowledge, we incorporated the epidemiological principles, formal causal frameworks, statistical theory and machine learning theory with the hope that it will foster rigor in future clinical studies using EHRs.

There were several notable limitations of this study. We used non-negative least square estimation as the meta-learning algorithm for the SuperLearner, although it is possible to use other machine learning classifiers including regularization methods, other ensemble-based trees or a neural network (Boehmke and Greenwell, 2019; Rose and Rizopoulos, 2020). The causal estimands of diabetes care provisions were generated using mainstream statistical algorithms in R software (v.4.1.0) which did not support the functionality to account for clustering arising due to repeated-measures outcomes. We may further diversify the collection of base learners with other machine learning classifiers including support vector machines, generalized additive models, multivariate additive regression splines (Rose, 2013). In this longitudinal design, we approximated the causal effects using the discretized (annual) time intervals rather than conceptualizing the causal effects under the framework of continuous-time. Although the estimation of causal effects using discrete time-intervals has been the standard practice in causal literature (Robins et al., 2000), the emerging research indicate how the inverse probability estimation using the continuous-time may produce statistical inference with desirable properties (e.g. more accuracy (i.e. reduced biased) and more precision (i.e. reduced standard errors) of the causal estimands) (Xiao et al., 2010). We characterized the individual-level causal estimation using the conditional average treatment effect rather than using the Bayesian non-parameteric formulation for estimating individualized treatment-response curves (Xu et al., 2016). In addition, the implementation of machine learning algorithms are often considered as "*black box*" due to their complexity. We may benefit from the incorporation of several recent advancements in machine learning discipline for generating longitudinal causal inference, and notable of which includes automated machine learning and interpretable machine learning (Boehmke and Greenwell, 2019). Future EHR studies may focus on the use of double machine learning algorithms for causal estimation in which regularization bias may be corrected using orthogonalization while overfitting bias may be corrected using cross-fitting (Chernozhukov et al., 2018). Future work may also focus on the use of targeted maximum likelihood estimation in which the marginal causal estimator is locally efficient with the smallest standard error and is robust to misspecification of either the treatment or the outcome model (Van der Laan and Rose, 2011). As an extension, it might be appropriate to construct confidence intervals of causal estimands using targeted bootstrap which is known to be robust to model misspecification and satisfy the regularity conditions of ensemble learning (Van der Laan and Rose, 2018).

Extra caution is necessary when drawing causal estimation using EHRs so that we can build trust as the mainstream statistical learning algorithms become tailored for EHR data in future (Kohane

et al., 2021). In conclusion, the use of causal prediction tools require a careful consideration given the high stakes involved with insurmountable implications for policy-makers. As eloquently stated by Nicholas Jewell, *"behind every data point there is a human story, there is a family, and there is suffering"* and thus it is necessary to engage in the complexities of EHRs when drawing causal estimation using machine learning algorithms (Rose and Rizopoulos, 2020). As a clinical utility, we hope that this study will facilitate discussions around the prevention of adverse chronic outcomes associated with diabetes through the improvement of diabetes care provision in primary care.

4.7 Funding acknowledgement

We like to acknowledge the SOPR-CIHR funding for NDR, and AHRQ Inspire PHC award for the application of this study. This statistical research was supported by Natural Sciences and Engineering Research Council (NSERC) PhD scholarship (CGS: 534600).

4.8 Acronyms

- AUROC = Area Under the Receiver Operating Curve
- ATC = Anatomical Therapeutic Chemical Classification System
- CPP = Cumulative Patient Profile
- CV= Cross-Validation
- EHR = Electronic Health Records
- EON = Easter Ontario Network
- i.i.d. = Independent and Identically Distributed
- ICD-9 = International Classification of Disease Version 9
- MSE = Mean Square Error
- NDR = National Diabetes Repository
- NNLS = Non-Negative Least Squares
- OHIP= Ontario Health Insurance
- RE = Relative Efficiency
- REB = Research Ethics Board
- SGLT-2i= Sodium Glucose Co-transporter 2 Inhibitors
- SMD = Standardized Mean Difference
- UTOPIAN = University of Toronto Practice-Based Research Network

4.9 Contributions

SK drafted this manuscript, carried out the analysis and produced the graphs. MG, TC and SK contributed to the conception of the research question. TC contributed to the data curation of this project. OS, ME, RM, JG, CM, ES and SK contributed to the study design, literature review and revisions to the manuscript. MG, BO, BA, FS, CP and SK contributed towards the clinical application and clinical relevance of this project. All authors critically reviewed the final paper.

4.10 Supplementary material

The section contains background information on base learners, along with supplementary results.

4.10.1 Base learners

We provide a brief summary of base learners including regularization methods, ensemble-based trees, and neural network.

Regularization methods

The Least Absolute Shrinkage and Selection Operator (LASSO), ridge regression and elastic net use the penalization terms $\lambda^T = \{\lambda_1, \lambda_2\}$ as the hyper-parameters to control for bias-variance trade-offs. The LASSO regression relies on L^1 penalization, ridge regression relies on L^2 penalization and elastic net relies on a combination of L^1 and L^2 penalization. We may summarize the objective function of LASSO, ridge and elastic net regressions with respect to the penalization terms λ^T as:

$$\text{Regularization models} = \begin{cases} \text{LASSO:} & min\left(SSE + \lambda_1 \sum_{k=1}^{p} |\beta_j|\right) \\ \text{Ridge:} & min\left(SSE + \lambda_2 \sum_{k=1}^{p} \beta_j^2\right) \\ \text{Elastic Net:} & min\left(SSE + \lambda_1 \sum_{k=1}^{p} |\beta_j| + \lambda_2 \sum_{k=1}^{p} \beta_j^2\right) \end{cases} . \tag{4.11}$$

where SSE denotes the squared sum of errors and the hyper-parameters λ_1 and λ_2 characterize the L^1 and L^2 penalization terms, respectively. An increase in penalty terms $\lambda^T = \{\lambda_1, \lambda_2\}$ reduce the magnitude of regression coefficients, and thereby introducing more bias at the expense of reducing variability in estimation of the regularization model (Rubin and van der Laan, 2006). The regularization models are tuned using the complexity parameter λ^T in which $k-$fold cross-validation is performed in the training set to minimize the mean square error (defined as the sum of bias squared and variance).

Ensemble-based trees

As an alternative to parametric assumption (e.g. linearity) with regularization models, we may consider non-parametric ensemble-based trees (e.g. bagging, random forest, gradient boosting) for hypothetical estimation of diabetes provision in future. A single classification tree relies on recursive partitioning of the predictor space to minimize the mean square error. An ensemble of trees are used to generate estimation based on the most common class of discrete outcome in the terminal node (i.e. node purity using the Gini index or cross-entropy). Bagging generates estimation using an ensemble of trees in which

bootstrap aggregation (sampling with replacement) is applied on each classification tree. Unlike bagging, random forest provides improvement in estimation by de-correlating the trees using a random selection of $\sqrt{p}$ predictors. Both bagging and random forest algorithms generate estimation using a collection of independent trees. In contrast, the classification trees are grown sequentially using the gradient boosting algorithm in which the knowledge to improve estimation is propagated sequentially across trees using the learning parameter η (Boehmke and Greenwell, 2019).

Support vector machines

Support vector machines use the "*kernel function*" to enlarge the feature space in higher dimensions so that the hyperplanes can be used to distinguish between two classes: (i) patients with optimal diabetes provision; (ii) patients with sub-optimal diabetes provision. Support vector machines allow the hyperplanes to be constructed based on the hard margin classifier or the soft margin classifier. The hard-margin classifier may be described as an infinite number of separating hyperplanes in which the two classes can be perfectly separated. The decision boundary based on hard-margin classifier is constructed by maximizing the Euclidean distance between the two classes. In contrast, soft-margin classifier ignores the perfect separation of two classes (even if it is achievable). The decision boundaries based on the soft-margin classifier uses the slack parameter ξ to allow some data points to be on the wrong side of the margin. Although counter-intuitive, the soft-margin classifiers may provide more reliable causal predictions with improved model stability as compared to hard-margin classifiers in the presence of outliers (Boehmke and Greenwell, 2019).

Neural network

Most machine learning algorithms are shallow in a sense that few layers of data transformations are imposed to generate hypothetical estimation. However, the use of shallow machine learning algorithms for the longitudinal diabetes cohort may not be appropriate due to large dimension of the predictor space (e.g. diagnostic ICD9 codes, billing OHIP codes, ATC codes). As an alternative, deep hypothetical estimation in the context of multi-layer neural network may be ideal to generate prediction by mapping a large dimension of input features to the target output (e.g. "diabetes provision") using the appropriate data transformation and feedback signals (Lim et al., 2018).

We may describe the neural network architectures using three layers: (i) input layer, (ii) hidden layers, (iii) output layer. The input layer incorporates the features in the neural network architecture while the hidden layer transforms the inputs to learn different attributes and to generate a signal for "diabetes care provisions" in the output layer. We build the neural network using the following configuration of the hyperparameters: (i) number of units in the hidden layer, (ii) weight decay, (iii) activation function. The number of units in the hidden layer and weight decay are the building blocks of neural network architectures and we may use them to gauge the complexity of the neural network in terms of memorization capacity. The activation functions (e.g. linear, logistic) are mathematical expressions used to determine if there is enough informative input to fire a signal to the next layer. In large-scale databases, the neural network architecture may rely on *"mini-batch stochastic gradient descent"* in which the back-propagation algorithm is used to identify the optimum of the objective function (Boehmke and Greenwell, 2019).

4.10.2 Supplementary results

The supplementary results include descriptive tables for diabetes care provision in 2018 and 2019. A summary of covariate balance is provided using the standardized mean difference as the evaluation metric.

Table 4.3: Diabetes provision in 2018 using 2017 predictors in Diabetes Action Canada Repository

			Diabetes provision				
Characteristics			**Less than adequate care**		**Adequate to optimal care**		**Total**
Age groups			**N**	**Row Percent (%)**	**N**	**Row Percent (%)**	**N**
40-49 years			1,386	34.9%	2,587	65.1%	3,973
50-59 years			2,553	26.6%	7,043	73.4%	9,596
60-69 years			2,502	19.4%	10,423	80.6%	12,925
70-79 years			1,843	15.9%	9,734	84.1%	11,577
80+ years			1,624	24.1%	5,107	75.9%	6,731
Sex							
Female			5,158	23.8%	16,474	76.2%	21,632
Male			4,750	20.5%	18,420	79.5%	23,170
Income Quintiles							
1			2,200	20.9%	8,303	79.1%	10,503
2			1,878	21.6%	6,818	78.4%	8,696
3			1,746	22.4%	6,041	77.6%	7,787
4			1,629	22.6%	5,588	77.4%	7,217
5			1,935	22.7%	6,590	77.3%	8,525
Missing			520	25.1%	1,554	74.9%	2,074
Rurality							
0			8,492	22.1%	29,868	77.9%	38,360
1			1,416	22.0%	5,026	78.0%	6,442
Metformin	**Sulfonylurea**	**SGLT-2i**					
0	0	0	6,763	31.3%	14,815	68.7%	21,578
0	0	1	86	16.0%	450	84.0%	536
0	1	0	216	16.5%	1,093	83.5%	1,309
0	1	1	25	11.7%	188	88.3%	213
1	0	0	2,036	14.3%	12,172	85.7%	14,208
1	0	1	200	12.1%	1,454	87.9%	1,654
1	1	0	485	11.4%	3,770	88.6%	4,255
1	1	1	97	9.2%	952	90.8%	1,049
Total			9,908	22.1%	34,894	77.9%	44,802

Table 4.4: Diabetes provision in 2019 using 2018 predictors in Diabetes Action Canada Repository

			Diabetes provision				
Characteristics			**Less than adequate care**		**Adequate to optimal care**		**Total**
Age groups			**N**	**Row Percent (%)**	**N**	**Row Percent (%)**	**N**
40-49 years			1,294	34.8%	2,423	65.2%	3,717
50-59 years			2,355	26.1%	6,661	73.9%	9,016
60-69 years			2,866	22.4%	9,924	77.6%	12,790
70-79 years			2,320	19.3%	9,683	80.7%	12,003
80+ years			1,936	27.4%	5,135	72.6%	7,071
Sex							
Female			5,571	26.0%	15,871	74.0%	21,442
Male			5,200	22.5%	17,955	77.5%	23,155
Income Quintiles							
1			2,398	23.0%	8,033	77.0%	10,431
2			1,988	23.2%	6,599	76.8%	8,587
3			1,940	25.0%	5,821	75.0%	7,761
4			1,822	25.2%	5,405	74.8%	7,227
5			2,130	24.8%	6,450	75.2%	8,580
Missing			493	24.5%	1,518	75.5%	2,011
Rurality							
0			9,184	24.1%	28,988	75.9%	38,172
1			1,587	24.7%	4,838	75.3%	6,425
Metformin	**Sulfonylurea**	**SGLT-2i**					
0	0	0	6,571	32.2%	13,817	67.8%	20,388
0	0	1	112	16.2%	581	83.8%	693
0	1	0	247	19.2%	1,037	80.8%	1,284
0	1	1	37	14.9%	211	85.1%	248
1	0	0	2,717	18.7%	11,834	81.3%	14,551
1	0	1	304	13.2%	1,997	86.8%	2,301
1	1	0	617	16.1%	3,205	83.9%	3,822
1	1	1	166	12.7%	1,144	87.3%	1,310
Total			10,771	24.2%	33,826	75.8%	44,597

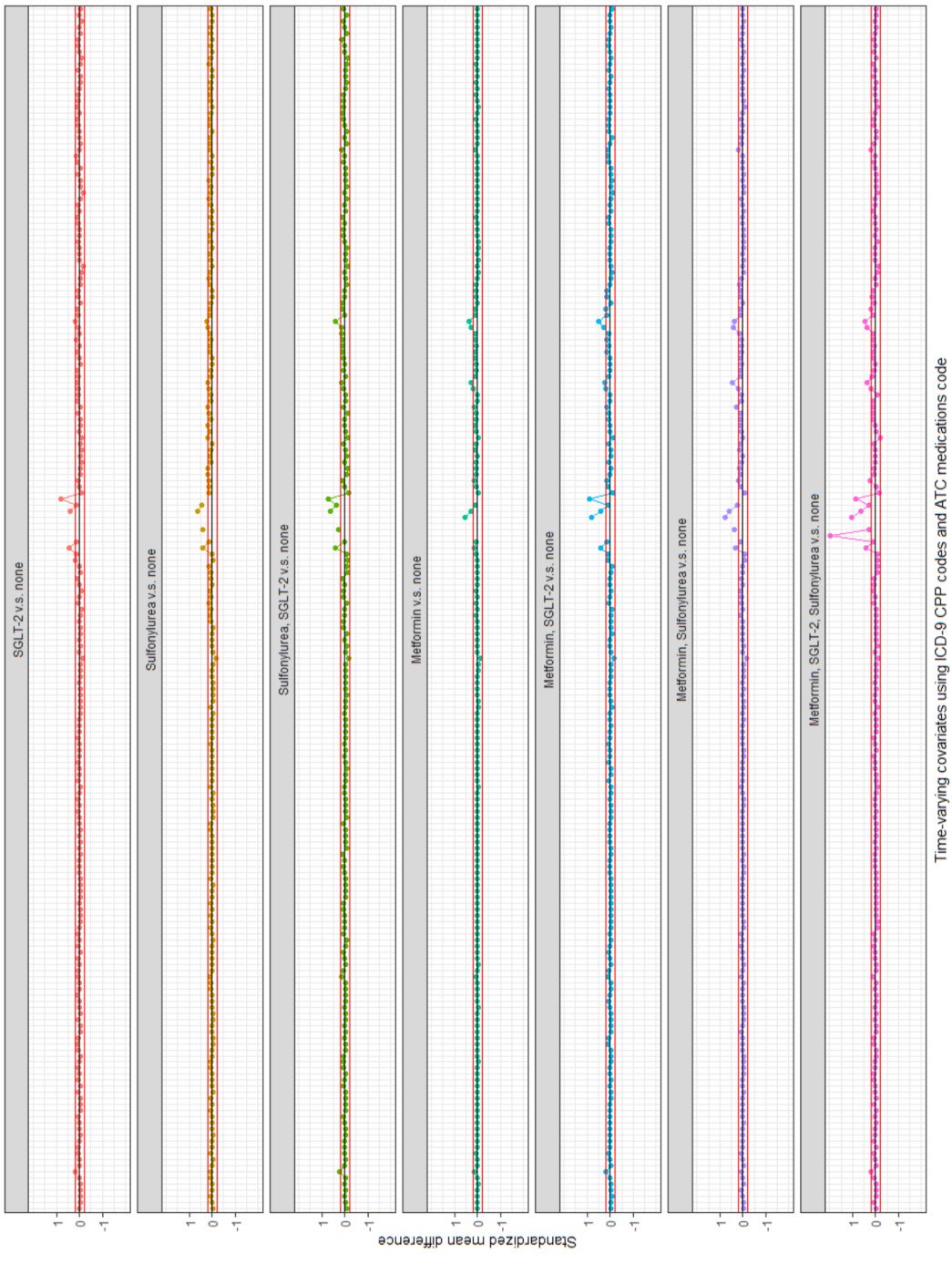


Figure 4.7: Covariate balance in longitudinal diabetes cohort using calibrated weights with respect to treatment naive cohort