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

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

***Abstract*— As different scientific disciplines begin to converge on machine learning for causal inference, we demonstrate the application of machine learning algorithms in the context of longitudinal causal estimation using electronic health records. Our aim is 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 [27] 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 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, and (vi) neural network. Each statistical learning algorithm was fitted using the pseudo-population generated from the marginalization of the time-dependent confounding process. 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 have improved 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 type II diabetes through the improvement of diabetes care provisions in primary care.**

***Index Terms*—Causal Inference, Machine Learning, SuperLearner, Longitudinal Interventions, Chronic Disease Prevention, Electronic Health Records, Primary Care**

I. Introduction

We may describe the multi-faceted data analytics landscape using three paradigms: (i) data exploration, (ii) inference and (iii) prediction. Causal methods focus on an inference paradigm in which hypothetical interventions are constructed, and the philosophical discussions around "causal methods" can be traced back many centuries [14]. In this article, our aim was to formulate marginal structural models in which we envisioned hypothetical (i.e. counterfactual) treatment regimes using several machine learning algorithms. In particular, we constructed the hypothetical treatment cohorts using a treatment naïve cohort and treatment drop-in cohort. We described the "treatment naïve" cohort as the absence of treatment regimen while the "treatment drop-in" cohort as the initiation of treatment post-baseline [22]. As an example, we considered a hypothetical cohort in which the type II diabetes patients were not prescribed glucose-lowering medications during the study period, and we use this cohort to describe the treatment naïve cohort.

It is essential to distinguish between the etiological and the intervening genres of causality in medicine [17]. In this article, we emphasized that the hypothetical treatment 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 intervened 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 [37]. 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 [35].

Our aim was to construct hypothetical estimation of diabetes care provisions (in future) by reducing bias arising due to temporal confounding and other epidemiological sources. For example, the use of older glucose-lowering medications (e.g. Sulfonylurea) might have been associated with worse health outcomes than newer glucose-lowering medications (e.g. SGLT-2i). We described 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 [37]. Although the randomization procedure in controlled experiments nullifies these causal challenges whereby the controlled experiments are by design unconfounded and associations imply causation [13], we need to account for these causal and statistical challenges when drawing valid estimation from longitudinal cohorts. This, in turn, allow us to generate reliable estimation with greater scope of generalizability when the application of machine learning algorithms is shifted from training sample to test or validation sample.

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

[‡] Corresponding author, Address: [i] Department of Family and Community Medicine, Temerty Faculty of Medicine, University of Toronto, 500 University Ave (Room 345), Toronto, ON M5G 1V7; [ii] Dalla Lana School of Public Health, 155 College St Room 500, Toronto, ON M5T 3M7 (e-mail: sumeet.kalia@utoronto.ca).

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).

A. Motivation and Knowledge gap

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 were not directly observable in practical sense, the aim of this study was to facilitate the discussion around the prevention of chronic adverse outcomes associated with diabetes through the improvement of diabetes care provision in primary care.

II. Materials and Methods

The material section described the data source, and the methods section is split into two sub-sections: (i) notational framework and (ii) machine learning algorithms. The notational framework described 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 were described so that we can construct the stacked estimation using the SuperLearner framework.

A. 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 curated EHRs on patients living with diabetes across multiple practice-based research networks (PBRNs) located in Alberta, Manitoba, Quebec, Ontario, and Newfoundland. As of July 1^{st} 2020, the NDR collected 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 contained patient-level demographics, medical diagnosis, procedures, medications, immunization, laboratory test results, vital signs and risk factors. Since the EHRs in NDR comprised 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 [38]. 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 [1]. Analyses were performed using the R software (v.4.1.0) in a Secure Analytic Virtual Environment at the Centre for Advanced Computing located at Queen's University, Ontario, Canada.

B. Notational framework

We specified the notational framework using the potential outcomes (i.e. counterfactual outcomes). We introduced the notation for longitudinal repeated-measures outcomes, followed by sequential variants of causal assumptions. We formulated a stabilizing weight function with calibrated restrictions to account for time-dependent confounding process.

1) 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 denoted 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 denoted the patient demographics with respect to k^{th} baseline covariates as X_{ik}. We partitioned 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 included 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 included vaccination, lifestyle information, annual laboratory requisition, billing ICD9 codes and Ontario Health Insurance (OHIP) billing codes. We denoted 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 constructed 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 $\bar{Y}_{ij} = \{Y_{i1}, Y_{i2}, ...Y_{ij}\}$. We described the latent health status for patients i as U_i. For the sake of brevity, we suppressed 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.

2) Diabetes care provision: We described *"diabetes care provisions"* using a modification of the summary quality index inspired by Grunfeld *et al.* [11] and Nietert *et al.* [27]. We defined the longitudinal primary endpoint for diabetes care provisions as the sum of eight elements as

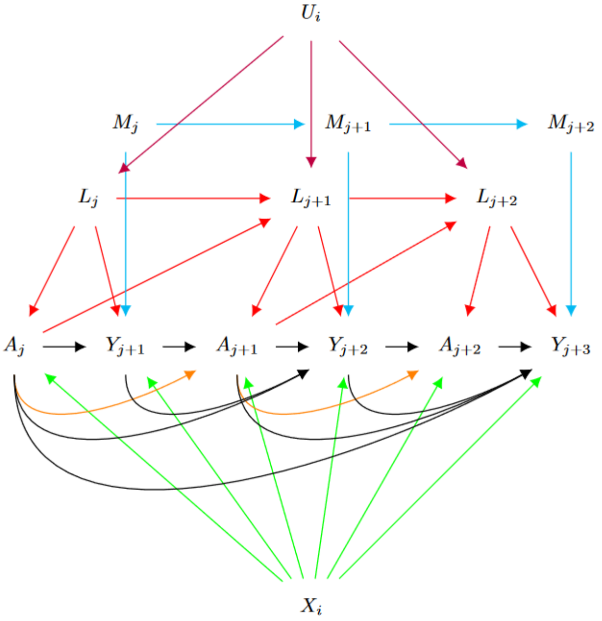
$$
\begin{aligned}
(\text{Diabetes care score})_j =& \mathbb{1}(\text{Visit count} \geq 2)_j \\
&+ \mathbb{1}(\text{Blood pressure count} \geq 2)_j \\
&+ \mathbb{1}(\text{Weight count} \geq 2)_j \\
&+ \mathbb{1}(\text{Hemoglobin A1c count} \geq 2)_j \\
&+ \mathbb{1}(\text{Lipid count} \geq 1)_j \\
&+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

where $\mathbb{1}(\cdot)$ denoted the indicator function indexed with respect to calendar year j. We further defined 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 binarized the longitudinal score of $(\text{Diabetes care score})_j$ as

$$
Y_{ij} = \begin{cases} 1 = \text{Adequate service: (Diabetes care score)}_j \in \{4, 5, 6, 7, 8\} \\ 0 = \text{Inadequate service: (Diabetes care score)}_j \in \{0, 1, 2, 3\} \end{cases} . \tag{1}
$$

3) Identifiability assumptions: Identifiability assumptions were necessary to ensure that we estimated the causal estimands from longitudinal studies with observational design. The necessary identifiability assumptions included: (i) sequential exchangeability; (ii) sequential postivity; (iii) sequential consistency [13]. We described the sequential exchangability as "*no unmeasured confounding*" whereby the probability of treatment assignment at each discretized time point j was independent of the potential outcome (with respect to the causal treatment regimes) conditioned on the observed history. We may write the sequential exchangability assumption as $Y_j^g \perp A_j | \bar{\mathcal{H}}_{j-1}$ where Y_j^g denoted the potential outcome under the causal treatment regime g, and where $\bar{\mathcal{H}}_{j-1} \equiv \{\bar{A}_{i,j-1}, \bar{L}_{i,j-1,k}, \bar{M}_{i,j-1,k}, \bar{Y}_{i,j-1}, X_{ik}\}$

Fig. 1. Directed acyclic graph with time-dependent treatment-confounder feedback



was the observed history up to and including time point $j-1$. We described 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 write the sequential positivity assumption as $P(A_j|\bar{\mathcal{H}}_{j-1}) > 0$. The sequential consistency assumption was 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 write the sequential consistency assumption as $Y^g_j = Y^{\bar{a}}_j$ where $g = \bar{a}$. We used the potential framework to formulate the causal models for $Y^{\bar{a}}_j$ in which we estimate the diabetes care provisions with respect to causal interventions $\bar{a}$. We assumed that the censoring mechanism C_{ij} was completely at random in which the censoring process was independent of discretized time points T_{ij} and longitudinal outcome Y_{ij}, conditioned on observed cumulative history $\bar{\mathcal{H}}_{ij}$ as $C_{ij} \perp \{Y_{ij}, T_{ij}\}|\bar{\mathcal{H}}_{i,j-1}$.

4) Model-based dynamic estimation: We used the directed acyclic graph (Figure 1) to describe the relationships among time-dependent treatment process A_j, time-varying covariates L_{jk}, time-varying outcome-predictors M_{jk}, baseline covariates X_{jk}, latent health status U_i, and repeated-measures outcome Y_j. We used the directed acyclic graph to describe the treatment-confounder feedback, denoted using red edges in Figure (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) [25]. In this article, we described 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). We conceptualized the time-dependent confounders L_{jk} as a surrogate measure to capture the latent health status of patients.

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

$$\Psi^{\bar{a}}_{ij} = Pr(Y^{\bar{a}}_{ij}|\bar{A}_{i,j-1}, \bar{M}_{i,j-1}) = \Phi\left(\bar{a}_{ij}, m_{i,j-1,k}\right) \qquad (2)$$

where $\Phi(\cdot)$ denoted an arbitrary marginal function of outcome process with respect to time-dependent covariates $L_{i,j-1,k}$ and Y_{ij-1}. We noted the exclusion of time-dependent confounders in Equation (2) because this may bias the direct or indirect treatment effects in the longitudinal causal structure [31]. In similar fashion, we encoded the treatment model with respect to time-dependent covariate process as

$$Pr(A_{ij}|\bar{\mathcal{H}}_{i,j-1}/\bar{M}_{i,j-1}) = \Omega\left(x_i, l_{ijk}, y_{ij}, a_{i,j-1}\right) \qquad (3)$$

where $\Omega(\cdot)$ denoted an 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.

5) Dynamic estimands using causal treatment modalities: We evaluated 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 formalized the pairwise estimands for hypothetical treatment modality $\bar{a}$ under the dynamic treatment regimen as

$$\text{Average treatment effect} = \left(\Psi^{\bar{a}}_{ij} - \Psi^{\bar{a}'}_{ij}\right) \times 100\% \qquad (4)$$

where $\Psi^{\bar{a}}_{ij}$ characterized the hypothetical outcome probability with respect to treatment modality $\bar{a}$, and where $\bar{a} \neq \bar{a}'$. We formulated 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 were indexed with respect to longitudinal calendar year (i.e. $j = \{2016, 2017, 2018\}$), this gave rise to $(2^3)^3 = 512$ possible treatment regimen. We restricted 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 led 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).

6) Stabilizing weight function: We introduced the stabilizing weight function to reduce 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 described the stabilizing weight function with respect to longitudinal treatment process A_{ij} as

$$SW^{\bar{A}}_{ij} = \prod_{t=1}^{j} \frac{Pr(A_{it}|\bar{\mathcal{H}}_{i,t-1}/\{L_{i,j-1,k}, Y_{ij-1}\})}{Pr(A_{it}|\bar{\mathcal{H}}_{i,t-1})} \qquad (5)$$

where the numerator $Pr(A_{ij}|\bar{\mathcal{H}}_{i,j-1}/\{L_{i,j-1,k}, Y_{ij-1}\})$ described the stabilizing factor with the exclusion of time-dependent covariates while the denominator $Pr(A_{ij}|\bar{\mathcal{H}}_{i,j-1}) \equiv Pr(A_{ij}|L_{i,j-1}, Y_{ij-1})$ described the inverse probability of treatment assignment with the inclusion of time-dependent covariates. Pajouheshnia *et al.* [28] 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 considered 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. Instead, the stabilized inverse probability treatment weights (with the calibrated restrictions) were used to create the pseudo-population in which the time-dependent treatment process A_{ij} was exogenous. Similar to Dong [9], we truncated 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 [47].

7) Calibration of stabilizing weight function: In survey sampling, the calibration of weight functions is performed to integrate the auxiliary information in which the distance between the initial weights and final weights is minimized subject to calibrating restrictions [8]. We introduced the calibration framework in this article to improve the finite-sample covariate balance of the stabilizing weight function [48]. In particular, we formulated 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

$$\sum_{i=1}^{n}\sum_{j=2016}^{2019} SW_{ij}^{\bar{A}}(\lambda)\sum_{t=1}^{j}\Big[(A_{it}-\hat{e}_{it}^{A})\times L_{ik,t-1}\Big]=0 \qquad (6)$$

where $SW_{ij}^{\bar{A}}(\lambda) = SW_{ij}^{\bar{A}} \times exp(K\lambda)$ denoted the calibrated stabilized weights with the unknown parameter λ and data-dependent covariate restrictions in matrix K. In Equation (6), we observed 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 $SW_{ij}^{\bar{A}}(\lambda)$ were constrained to be non-negative. This orthogonality constraint ensured that the propensity score residuals are linearly independent with respect to the time-varying covariates $L_{ik,t-1}$ in high-dimensional Euclidean space [32].

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) [12], this property is not guaranteed to hold for calibrated stabilized weights [48]. In additional to the time-dependent covariate balancing constraints (above in Equation (6)), we also imposed the restriction for average calibrated weights to be equal to one at each time-point j as

$$E(SW_{ij}^{\bar{A}}(\lambda)) = \frac{1}{n}\sum_{i=1}^{n} SW_{ij}^{\bar{A}}(\lambda) = 1. \qquad (7)$$

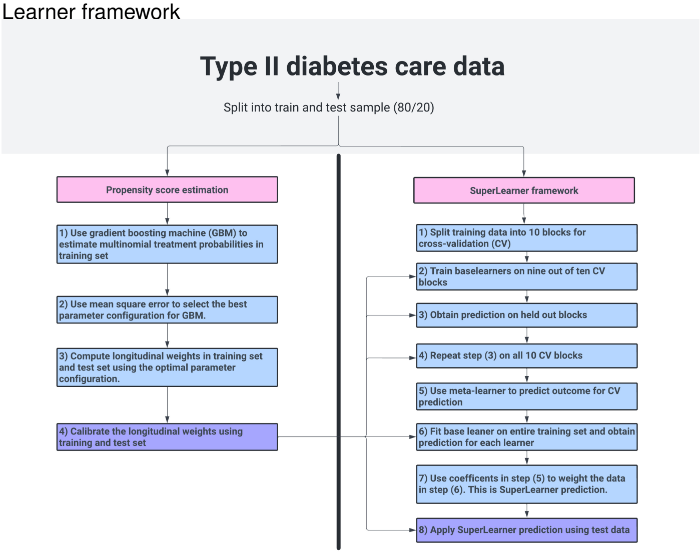
We used the calibrated weights satisfying Equation (6) and (7) to construct the pseudo-population for the longitudinal diabetes cohort and to assess the covariate balance in hypothetical treatment regimes with respect to metformin, sulfonylurea and SGLT-2i. The constrained optimization was implemented using the Barzilai-Borwein gradient method in R software [43].

C. Machine learning algorithms

In similar spirit to Blakely *et al.* [5] and Karim *et al.* [16], our aim was to estimate the marginal means using the machine learning algorithms. We were interested in conducting supervised machine learning using a collection of mainstream statistical learning algorithms including least absolute shrinkage and selection operator (lasso), ridge regression, elastic net, random forest, gradient boosting machine and neural network. We provided a brief summary of each base learner in the Supplementary Section.

1) SuperLearner: The SuperLearner algorithm combined the estimation from individual base learner to create a stacked estimation [7]. Since both causal effects and longitudinal estimation (in the context of machine learning) can be described as an estimation problem, the idea was to further improve the causal estimation using the SuperLearner in which the stacked estimand was indexed with respect to multiple base learners [41]. In earlier settings, the SuperLearner algorithm outperformed individual base learners (e.g. regularization methods, ensemble-based trees or deep learning using neural networks) to generate an optimal system for estimation [40].

Fig. 2. Analytic workflow for propensity score estimation and SuperLearner framework


Type II diabetes care data
Split into train and test sample (80/20)
Propensity score estimation
1) Use gradient boosting machine (GBM) to estimate multinomial treatment probabilities in training set
2) Use mean square error to select the best parameter configuration for GBM.
3) Compute longitudinal weights in training set and test set using the optimal parameter configuration.
4) Calibrate the longitudinal weights using training and test set
SuperLearner framework
1) Split training data into 10 blocks for cross-validation (CV)
2) Train baselearners on nine out of ten CV blocks
3) Obtain prediction on held out blocks
4) Repeat step (3) on all 10 CV blocks
5) Use meta-learner to predict outcome for CV prediction
6) Fit base leaner on entire training set and obtain prediction for each learner
7) Use coefficents in step (5) to weight the data in step (6). This is SuperLearner prediction.
8) Apply SuperLearner prediction using test data

Unlike 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 [6]. In similar spirit to Rose [34], we formulated the SuperLearner algorithm in the context of hypothetical estimation using the following steps:

1) We selected the brute-force configuration of the hyperparameter grid search (see Supplementary section) 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 applied the patient-level data split on training sample to create 10 mutually exclusive and exhaustive blocks of equal (or approximately equal) size. We applied the clustered 10-fold cross-validation (CV) in which the cumulative-time product treatment weights were preserved for each patient within 10 blocks.
3) We fitted each machine learning algorithm (i)-(vi) using 10-fold CV with calibrated weights. We used the validation set in the training sample (using 10-fold CV) to predict the probability of diabetes provision $\Psi_{ij}^{\bar{a}}(W)$ for i^{th} individual at j^{th} time-point for w^{th} machine learning algorithm.
4) We collected the estimated probabilities $\Psi_{ij}^{\bar{a}}(W)$ for the entire training set and then estimate the CV MSE for each machine learning algorithm w (see Equation (10)).
5) We estimated the optimal weight combinations for machine learning algorithms indexed with respect to the weight vector α using the non-negative least square estimation as
$$\Psi_{ij}^{\bar{a}}(SL) = \sum_{l=1}^{L}\alpha_l \Psi_{ij}^{\bar{a}}(W)$$
where α_l were the SuperLearner weights and $\Psi_{ij}^{\bar{a}}(SL)$ denoted the predicted probability of the SuperLearner.
6) We used 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 performed sample-split on each independent patient units [3].

Splitting the training and test set at the patient level (rather than at the repeat observations) preserved the cumulative-time products of stabilized weight function within each sample split, and reduced the time-dependent confounding process. We estimated the counterfactual probabilities (in the test sample) with respect to eight treatment groups (separately) for each base learner with non-negative weight contributions to the SuperLearner. The counterfactual probabilities of the base learners were then amalgamated using the non-negative least squares to generate stacked estimations for each counterfactual treatment.

D. Implementation of Machine Learning pipelines

We described 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 described 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.

1) Generation of longitudinal diabetes cohort: We constructed a longitudinal diabetes cohort in which patients were enrolled when the following conditions were satisfied: (i) patients were at least 40 years of age as of January 1^{st} of each index year; (ii) patient had an indication in EHRs corresponding to diabetes; (iii) research quality criteria for EHRs was satisfied [39]; (iv) patient had at least one visit recorded in billing or encounter fields within calendar year; (v) type I diabetes patients were excluded [44]. The age restriction for condition (i) was in agreement with the diabetes provision guidelines [30], while condition (ii) was borrowed from earlier work on diabetes phenotype [45]. We imposed administrative censoring where the patients were censored at the end of the study period (December 31, 2019). We used the open cohort design with time-dependent risk-set to make hypothetical estimation of diabetes care provision. We enriched the prediction matrix with elements captured from EHRs 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.

2) Data splitting: During the data pre-processing step, it was necessary to prevent *"data leakage"* whereby the information may propagate outside the training set [18]. 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 mitigated the possibility of *"data leakage"* with two data pre-processing steps. First, we generated a dynamic cohort in which the predictors (including the individual elements of diabetes care) were time-lagged with one calendar year with respect to the composite binary outcome of "diabetes provision". Second, we performed the data splitting step for training sample and testing sample prior to re-sampling iterations of machine learning algorithms. The second step ensured that we did not screen for any strong predictors prior to 10-fold CV [10]. Using the total number of unique patients as the sampling unit, we partitioned the longitudinal diabetes cohort data as 80% training sample and 20% test sample.

3) Feature Engineering: We captured several elements of primary care EHRs, and incorporated them as high-dimensional prediction matrix using *"one-hot"* (dummy) encoding. In particular, we implemented the feature engineering as boolean design matrix for the following elements in EHRs 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, SGLT-2i; (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 was implemented using binary encoding scheme and can be described as

$$\text{Feature}(t) = \begin{cases} 1 & \text{if present within calendar year } t \\ 0 & \text{if absent within calendar year } t \end{cases} \tag{8}$$

where we indexed each feature with respect to discrete calendar year t. We constructed 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 remained unchanged with respect to each index year from 2016 to 2019.

4) Marginalization of the covariate process: We applied 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 reduced 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 (II-B.6). The marginalization with respect to covariate process generated the hypothetical estimation for diabetes care provision. McCaffrey *et al.* [23] estimated propensity score for multiple treatment assignment using the generalized boosted models. Building on McCaffrey *et al.* [23], 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 built the stabilized weight functions as discrete cumulative-time product to account for the time-dependent confounding and then used the calibrated constraints to improve covariate balance in the pseudo-population of longitudinal diabetes cohort (as described in Section (II-B.6)).

5) Tuning hyperparameter grid search: We constructed a hyperparameter grid for each machine learning algorithm using the factorial configuration (described in Supplementary section). We applied the hyperparameter grid of gradient boosting machine on the treatment process (i.e. glucose lowering medications) to compute the cumulative-time product weights. We applied the criteria for the minimization of MSE to achieve improved estimation of multinomial propensities of glucose-lowering treatment assignment, which were then transformed into cumulative-time product weight functions.

Once the calibrated weights were estimated, we used the hyperparameter grid of statistical learning algorithms to generate stacked estimation. In particular, we applied the hyperparameter grid of base learner to the training (and held-out 10-fold CV) set using the cumulative-time product weights. We stacked the CV prediction in the training set and externally validate the performance of the SuperLearner using the test set.

6) Standardized mean difference: We evaluated the covariate balance in the pseudo-population based on standardized mean difference (SMD). The covariate balance was assessed for k time-dependent covariates used in the treatment model as

$$\text{SMD}_{jk} = \frac{\hat{p}_t - \hat{p}_c}{\sqrt{\frac{(\hat{p}_t)(1-\hat{p}_t)+(\hat{p}_c)(1-\hat{p}_c)}{2}}} \tag{9}$$

where $\hat{p}_t$ denoted the weighted average of treatment drop-in cohorts while $\hat{p}_c$ denoted the weighted average for treatment naïve cohort. The denominator in equation (9) corresponded to pooled standard deviation of treatment and control regimen. The covariate balance in the pseudo-population was assessed using the difference in prevalence measured relative to the units of the pooled standard deviation [2].

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

$$MSE(w) = \frac{\sum_{i=1}^{n}\sum_{j=2016}^{2019}(Y_{ij} - \Psi_{ij}(W))^2}{N} \tag{10}$$

where Y_{ij} denoted the diabetes provision for individual i at j^{th} time-point, and N denoted the sample size of training set.

III. Results

We described 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 in the test sample using the SuperLearner.

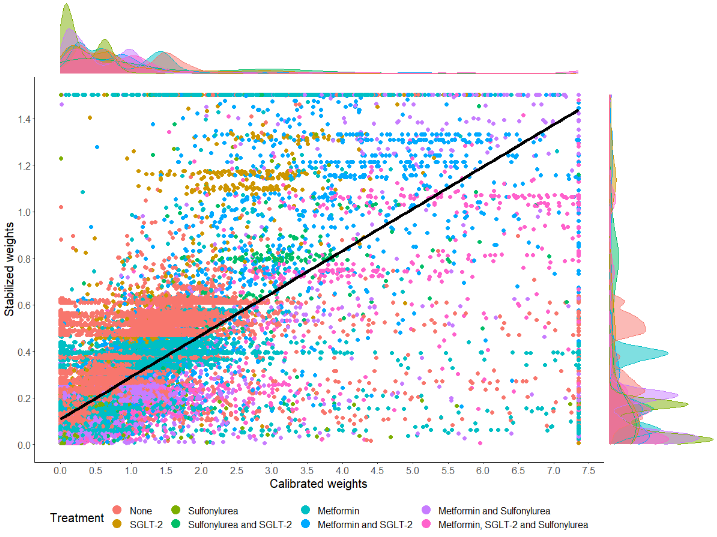
A. Cohort description

We noticed an improvement in diabetes provision with respect to increase in age groups with the exception for 80+ years (see Supplementary tables). Male patients tended to receive improved diabetes care with higher prevalence than female patients. A slight increase in prevalence of diabetes provision was observed in lowest income quintiles while no difference in prevalence of diabetes provision was observed 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%).

B. Covariate balance

The stabilized weight function was used to construct a pseudo-population in which the balance was achieved with respect to the distributions of the time-dependent covariates in each treatment regimen. Figure (3) describes the scatter plot between stabilized weights and calibrated weights for eight treatment groups. The side panels in Figure (3) show the density plots of stabilized and calibrated weights with respect to each treatment group. 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 was used to describe the covariate balance in each treatment cohort (using a combination of Metformin, Sulfonylurea and SGLT-2i) with respect to the treatment naïve cohort (i.e. no treatment regimen). 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 of SMD (see Supplementary Section).

Fig. 3. Scatterplot of stabilized and calibrated treatment weights



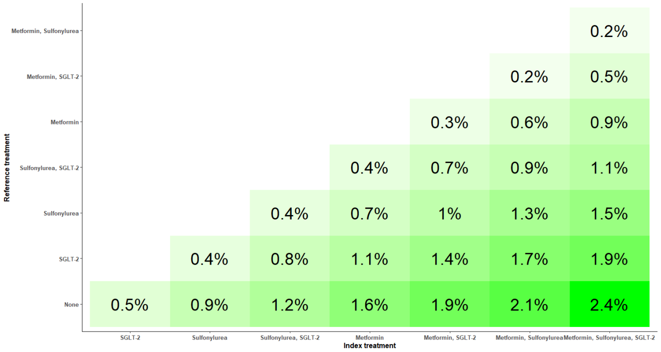
C. Stacked estimation using the SuperLearner algorithm

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 AUROC estimates of SuperLearner algorithm were higher than the AUROC of all base learners (see Supplementary Section).

D. Causal estimation

We generated the causal estimation with respect to homogeneous treatment groups in 2017, 2018 and 2019. Figure (4) describes the average treatment effect using causal risk difference between two mutually exclusive treatment groups 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 groups. As an example, the treatment group of metformin in each calendar year (i.e. 2016, 2017 and 2018) improved diabetes care provisions by 1.6% in relation to the treatment naïve cohort.

Fig. 4. Hypothetical risk difference using the SuperLearner prediction in test sample



IV. Discussion

There is a rich history for the application of statistical learning algorithms in the context of clinical epidemiology research of diabetes [4]. For example, the machine learning algorithms have been used to forecast glycaemia in Type 1 Diabetes Mellitus patients [33]. 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.

Kohane *et al.* [21] described 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 article 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.* [21], we described 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 [19] and we also strive for transparent data collection, and data harmonization procedures at NDR. 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 causal estimation due to the presence of data heterogeneity.

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 [22]. Balzer and Petersen [3] 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 groups and primary outcome. 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 1). Since time-dependent confounders existed 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 [46].

A. Limitations

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 [6], [35]. The causal estimands of diabetes care provisions were generated using 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 [34]. In this longitudinal design, we estimated 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 [31], the emerging research indicated 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) [46].

The implementation of machine learning algorithms were often considered as "*black box*" due to their complexity. We may benefit from the incorporation of several recent advancements in machine learning for generating longitudinal causal inference, and notable of which includes automated machine learning and interpretable machine learning [6]. Targeted maximum likelihood estimation (TMLE) is robust to misspecification of either the treatment or the outcome model [29], and TMLE may be applied to the longitudinal cohort to ensure proper standard errors in a case when either the treatment or outcome model is misspecified [24], [41]. It might be appropriate to construct confidence intervals of causal estimands using targeted bootstrap (or its bias-corrected analogue) which is known to be robust to model misspecification and also satisfy the regularity conditions of ensemble learning [42]. As an extension to this work, high-dimensional propensity score (HDPS) algorithm can be applied in this longitudinal cohort when predicting diabetes care provisions with time-varying treatment and confounders [26]. An appealing feature of HDPS algorithm includes an improvement in the performance of causal estimation through proxy adjustment of unmeasured confounders [36].

B. Conclusion

This study demonstrated that three common classes of diabetes medications (SGLT-2i, Meformin, Sulfonylurea) may improve the quality of diabetes care with respect to the appropriate provision of primary care resources. Moreover, patients who received diabetes screening services in previous year were likely to receive improved diabetes provision in next calendar year. 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 [15]. 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 [20]. As a clinical utility, we hope that this study will facilitate discussions around the prevention of adverse chronic outcomes associated with type II diabetes through the improvement of diabetes care provision in primary care.

V. 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, CM, JG, 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. The corresponding author (SK) takes the responsibility for ensuring the accuracy of the results.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety." arxiv preprint. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Peter C Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107, 2009.

[3] Laura B Balzer and Maya L Petersen. Invited commentary: Machine learning in causal inference—how do i love thee? let me count the ways. *American Journal of Epidemiology*, 2021.

[4] Sanjay Basu, Karl T Johnson, and Seth A Berkowitz. Use of machine learning approaches in clinical epidemiological research of diabetes. *Current Diabetes Reports*, 20(12):1–19, 2020.

[5] Tony Blakely, John Lynch, Koen Simons, Rebecca Bentley, and Sherri Rose. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. *International journal of epidemiology*, 49(6):2058–2064, 2020.

[6] Brad Boehmke and Brandon M Greenwell. *Hands-on machine learning with R*. CRC Press, 2019.

[7] Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

[8] Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

[9] Yixing Dong. Continuous-time marginal structural models for adverse drug effects in pharmacoepidemiology. Master's thesis, University of Toronto (Canada), 2021.

[10] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[11] Eva Grunfeld, Donna Manca, Rahim Moineddin, Kevin E Thorpe, Jeffrey S Hoch, Denise Campbell-Scherer, Christopher Meaney, Jess Rogers, Jaclyn Beca, Paul Krueger, et al. Improving chronic disease prevention and screening in primary care: results of the better pragmatic cluster randomized controlled trial. *BMC family practice*, 14(1):1–12, 2013.

[12] Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

[13] Miguel A Hernán and James M Robins. *Causal inference*. Boca Raton: Chapman & Hall/CRC, forthcoming, 2019.

[14] David Hume. *A treatise of human nature*. John Noon, London, 1739.

[15] Noah M Ivers, Maggie Jiang, Javed Alloo, Alexander Singer, Daniel Ngui, Carolyn Gall Casey, and H Yu Catherine. Diabetes canada 2018 clinical practice guidelines: key messages for family physicians caring for patients living with type 2 diabetes. *Canadian Family Physician*, 65(1):14–24, 2019.

[16] Mohammad Ehsanul Karim, Robert W Platt, and BeAMS Study Group. Estimating inverse probability weights using super learner when weight-model specification is unknown in a marginal structural cox model context. *Statistics in medicine*, 36(13):2032–2047, 2017.

[17] Igor Karp and Olli S Miettinen. On the essentials of etiological research for preventive medicine. *European journal of epidemiology*, 29(7):455–457, 2014.

[18] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21, 2012.

[19] Karim Keshavjee, Vijaya Chevendra, Kenneth E Martin, David Jackson, Babak Aliarzadeh, Lorne Kinsella, Raymond Turcotte, Sarah Sabri, and Tao Chen. Design and testing of an architecture for a national primary care chronic disease surveillance network in canada. In *ITCH*, pages 341–345, 2011.

[20] Tara Kiran, Gray Moonen, Onil K Bhattacharyya, Payal Agarwal, Harpreet S Bajaj, James Kim, and Noah Ivers. Managing type 2 diabetes in primary care during covid-19. *Canadian Family Physician*, 66(10):745–747, 2020.

[21] Isaac S Kohane, Bruce J Aronow, Paul Avillach, Brett K Beaulieu-Jones, Riccardo Bellazzi, Robert L Bradford, Gabriel A Brat, Mario Cannataro, James J Cimino, Noelia García-Barrio, et al. What every reader should know about studies using electronic health record data but may be afraid to ask. *Journal of medical Internet research*, 23(3), 2021.

[22] Lijing Lin, Matthew Sperrin, David A Jenkins, Glen P Martin, and Niels Peek. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagnostic and prognostic research*, 5(1):1–16, 2021.

[23] Daniel F McCaffrey, Beth Ann Griffin, Daniel Almirall, Mary Ellen Slaughter, Rajeev Ramchand, and Lane F Burgette. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414, 2013.

[24] Kelly L Moore and Mark J van der Laan. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in medicine*, 28(1):39–64, 2009.

[25] Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. An introduction to g methods. *International journal of epidemiology*, 46(2):756–762, 2017.

[26] Romain Neugebauer, Julie A Schmittdiel, Zheng Zhu, Jeremy A Rassen, John D Seeger, and Sebastian Schneeweiss. High-dimensional propensity score algorithm in comparative effectiveness research with time-varying interventions. *Statistics in medicine*, 34(5):753–781, 2015.

[27] Paul J Nietert, Andrea M Wessell, Ruth G Jenkins, Chris Feifer, Lynne S Nemeth, and Steven M Ornstein. Using a summary measure for multiple quality indicators in primary care: the summary quality index (squid). *Implementation Science*, 2(1):1–12, 2007.

[28] Romin Pajouheshnia, Noah A Schuster, Rolf HH Groenwold, Frans H Rutten, Karel GM Moons, and Linda M Peelen. Accounting for time-dependent treatment use when developing a prognostic model from observational data: A review of methods. *Statistica Neerlandica*, 74(1):38–51, 2020.

[29] Menglan Pang, Tibor Schuster, Kristian B Filion, Maria Eberg, and Robert W Platt. Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology (Cambridge, Mass.)*, 27(4):570, 2016.

[30] Alanna V Rigobon, Sumeet Kalia, Jennica Nichols, Babak Aliarzadeh, Michelle Greiver, Rahim Moineddin, Frank Sullivan, and Catherine Yu. Impact of the diabetes canada guideline dissemination strategy on the prescription of vascular protective medications: a retrospective cohort study, 2010–2015. *Diabetes Care*, 42(1):148–156, 2019.

[31] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[32] Joseph Lee Rodgers, W Alan Nicewander, and Larry Toothaker. Linearly independent, orthogonal, and uncorrelated variables. *The American Statistician*, 38(2):133–134, 1984.

[33] Ignacio Rodríguez-Rodríguez, José-Víctor Rodríguez, Wai Lok Woo, Bo Wei, and Domingo-Javier Pardo-Quiles. A comparison of feature selection and forecasting machine learning algorithms for predicting glycaemia in type 1 diabetes mellitus. *Applied Sciences*, 11(4):1742, 2021.

[34] Sherri Rose. Mortality risk score prediction in an elderly population using machine learning. *American journal of epidemiology*, 177(5):443–452, 2013.

[35] Sherri Rose and Dimitris Rizopoulos. Machine learning for causal inference in biostatistics. *Biostatistics*, 21(2):336–338, 2020.

[36] Sebastian Schneeweiss. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clinical epidemiology*, 10:771, 2018.

[37] Uri Shalit. Can we learn individual-level treatment policies from clinical data? *Biostatistics*, 21(2):359–362, 2020.

[38] Xu Shi, Xiaoou Li, and Tianxi Cai. Spherical regression under mismatch corruption with application to automated knowledge translation. *Journal of the American Statistical Association*, pages 1–12, 2020.

[39] Karen Tu, Babak Aliarzadeh, Tao Chen, and Sumeet Kalia. University of toronto family medicine report: Caring for our diverse population, 2020. Accessed: 12-24-2020.

[40] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[41] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[42] Mark J Van der Laan and Sherri Rose. *Targeted learning in data science*. Springer, 2018.

[43] Ravi Varadhan and Paul Gilbert. Bb: An r package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of statistical software*, 32(1):1–26, 2009.

[44] Alanna Weisman, Karen Tu, Jacqueline Young, Matthew Kumar, Peter C Austin, Liisa Jaakkimainen, Lorraine Lipscombe, Ronnie Aronson, and Gillian L Booth. Validation of a type 1 diabetes algorithm using electronic medical records and administrative healthcare data to study the population incidence and prevalence of type 1 diabetes in ontario, canada. *BMJ Open Diabetes Research and Care*, 8(1):e001224, 2020.

[45] Tyler Williamson, Michael E Green, Richard Birtwhistle, Shahriar Khan, Stephanie Garies, Sabrina T Wong, Nandini Natarajan, Donna Manca, and Neil Drummond. Validating the 8 cpcssn case definitions for chronic disease surveillance in a primary care database of electronic health records. *The Annals of Family Medicine*, 12(4):367–372, 2014.

[46] Yongling Xiao, Michal Abrahamowicz, and Erica EM Moodie. Accuracy of conventional and marginal structural cox model estimators: a simulation study. *The international journal of biostatistics*, 6(2), 2010.

[47] Yongling Xiao, Erica EM Moodie, and Michal Abrahamowicz. Comparison of approaches to weight truncation for marginal structural cox models. *Epidemiologic Methods*, 2(1):1–20, 2013.

[48] Sean Yiu and Li Su. Joint calibrated estimation of inverse probability of treatment and censoring weights for marginal structural models. *Biometrics*, 2020.