Supplementary Material

Lauren D. Liao, Emilie Højbjerre-Frandsen, Alan E. Hubbard, and Alejandro Schuler*

Supplementary Material for the Article "Prognostic Adjustment with Efficient Estimators to Unbiasedly Leverage Historical Data in Randomized Trials"

Appendix A Expectation calculation when incorporating unobserved covariate

By definition, an unobserved covariate U is never seen in real data but we include such a variable in simulation. We aim to demonstrate that even if the outcome model can be learned perfectly from historical data, if there exists an unobserved shift between the historical and trial sample, then the learned historical outcome model (prognostic model) can never be equivalent to the trial outcome model. We explicitly write out the expectation of the outcome Y given treatment A, baseline covariates W, and data set indicator D using an unobserved covariate U.

$$E[Y|A,W,D=d] = \int E[Y|A,W,U=u,D=d]p(u|A,W,D=d)\mathrm{d}u$$

A shift in the distribution $p(u|A,W,D=1) \neq p(u|A,W,D=0)$ of the unobserved covariate U will generally result in unequal conditional expectations, i.e., $E[Y|A,W,D=1] \neq E[Y|A,W,D=0]$.

This shift is the basis of the simulation for Figure 3.B. In our simulation, we have $U|A,W,D=1 \sim \text{Unif}(0,1)$ and $U|A,W,D=0 \sim \text{Unif}(\underline{u},\overline{u})$ where the limits $underline(u), overline(u)$ increase or decrease past 0 or 1 depending on the desired magnitude of covariate shift. The "oracle" prognostic score is

Lauren D. Liao, Alan E. Hubbard, Alejandro Schuler, Division of Biostatistics, University of California, Berkeley, CA, USA
Emilie Højbjerre-Frandsen, Biostatistics Methods and Outreach, Novo Nordisk A/S, Denmark

given by $E[Y|A, W, D = 1]$, i.e. always integrating over the correct (trial) density $U|A, W, D = 1 \sim \text{Unif}(0, 1)$.

By framing shifts in the conditional mean as shifts in an unobserved covariate we can directly control the magnitude of the change instead of manually specifying different conditional mean functions.

Appendix B Discrete super learner specifications for simulation and case study.

Appendix B.1 Discrete super learner specifications

Machine learning is performed through discrete super learner that the targeted maximum likelihood estimator internally leverages. For simplicity, the prognostic model is built using the discrete super learner as well. A discrete super learner selects from a set of candidate models (i.e., the library) to obtain a single, best prediction model via cross-validation. In this section, we describe the exact tuning parameters and set up for the simulation and case study.

Appendix B.2 Simulation set up

Cross-validation: 5-fold cross-validation is used to select the best candidate learner in the library for historical sample size 1,000, and 10-fold cross-validation for historical sample size less than 1,000.
Cross-fit: 5-fold Cross-fitting is employed.
Discrete super learner library: Multivariate Adaptive Regression Splines with the highest interaction to be to the 3rd degree, linear regression, extreme gradient boosting with specifications: learning rate 0.1, tree depth 3, crossed with trees specified 25 to 500 by 25 increments. Cases with *fitted* prognostic score include an augmented library that includes candidate learners with prognostic score in addition.
Discrete super learner specifications: loss function is specified to be the mean square error loss.

Appendix B.3 Case study set up

Cross-validation: 20-fold cross-validation is used to select the best candidate learner in the library.
Cross-fit: 20-fold Cross-fitting is employed.
Discrete super learner library: Multivariate Adaptive Regression Splines with the highest interaction to be to the 3rd degree, logistic regression, extreme gradient boosting with specifications: learning rate 0.1, tree depths 3, 5, and 10, crossed with trees specified 25 to 500 by 25 increments, random forest with trees specified

25 to 500 by 25 increments, k-nearest neighbor of specification 3, 4, 5, 7, and 9 number of nearest neighbors, k. Cases with *fitted* prognostic score include an augmented library that includes candidate learners with prognostic score in addition to without.

Discrete super learner specifications: loss function is specified to be the mean log likelihood loss.

Selected prognostic model: The selected learner from 20-fold cross validation is a linear regression model for both the reanalyses with $n = 419$ and $n = 100$. Details are written in Appendix H.

Appendix C Simulation results for different data generation processes

Tab. 1: Mean of empirically estimated bias, variance, and standard errors of them for the targeted maximum likelihood estimator with or without prognostic score across different DGPs. For all the scenarios the conditional means are shared with the heterogeneous effect DGP, except for the constant effect DGP. Unless otherwise specified $(\tilde{n}, n) = (1000, 250)$.

Scenario	Estimator *prog.*	Bias	Var.	SE Bias	SE var.	RMSE	Power	Coverage
heterogeneous effect	TMLE *none*	-0.031	5.918	0.048	0.041	2.432	0.660	0.956
	TMLE *fitted*	-0.081	4.843	0.064	0.005	2.201	0.727	0.955
	TMLE *oracle*	-0.066	4.827	0.036	0.004	2.197	0.743	0.955
	linear *none*	-0.005	11.113	0.026	0.026	3.332	0.413	0.953
	linear *fitted*	-0.096	10.438	0.037	0.030	3.231	0.421	0.952
	linear *oracle*	-0.083	10.485	0.037	0.029	3.237	0.425	0.951
	unadjusted *none*	-0.015	10.373	0.009	0.011	3.219	0.439	0.951
heterogeneous effect *second specification*	TMLE *none*	0.052	0.929	−0.003	0.008	0.965	0.788	0.943
	TMLE *fitted*	0.056	0.886	0.005	0.006	0.942	0.811	0.947
	TMLE *oracle*	0.050	0.893	−0.006	0.006	0.946	0.813	0.943
	linear *none*	0.077	1.400	0.020	0.008	1.185	0.598	0.947
	linear *fitted*	0.016	0.980	0.025	0.003	0.990	0.740	0.951
	linear *oracle*	0.014	0.993	0.010	0.003	0.996	0.745	0.949
	unadjusted *none*	0.060	1.350	−0.002	0.010	1.163	0.615	0.952
constant effect	TMLE *none*	0.024	0.108	−0.001	0.000	0.330	0.656	0.945
	TMLE *fitted*	0.015	0.075	0.010	0.000	0.274	0.790	0.956
	TMLE *oracle*	0.011	0.073	0.004	0.000	0.271	0.816	0.958
	linear *none*	0.043	0.406	0.030	0.001	0.639	0.200	0.948
	linear *fitted*	0.015	0.070	0.012	0.000	0.265	0.814	0.964
	linear *oracle*	0.017	0.064	0.011	0.000	0.254	0.849	0.963
	unadjusted *none*	0.010	0.815	−0.015	0.001	0.903	0.153	0.941
small observed shift	TMLE *none*	-0.081	5.757	0.080	0.042	2.400	0.648	0.959

Continuation of Table 1

Scenario	Estimator *prog.*	Bias	Var.	SE Bias	SE var.	RMSE	Power	Coverage
	TMLE *fitted*	-0.133	5.513	0.058	0.039	2.350	0.686	0.950
	TMLE *oracle*	-0.098	4.918	0.014	0.004	2.219	0.738	0.959
	linear *none*	-0.068	11.084	0.027	0.028	3.328	0.409	0.951
	linear *fitted*	-0.061	11.147	0.023	0.028	3.338	0.407	0.950
	linear *oracle*	-0.125	10.429	0.041	0.031	3.230	0.426	0.947
	unadjusted *none*	-0.078	10.683	−0.043	0.011	3.268	0.439	0.946
small unobserved shift	TMLE *none*	-0.084	6.257	−0.028	0.039	2.502	0.642	0.948
	TMLE *fitted*	-0.121	5.761	0.001	0.015	2.402	0.661	0.949
	TMLE *oracle*	-0.102	4.989	−0.003	0.004	2.235	0.718	0.956
	linear *none*	-0.049	10.546	0.112	0.026	3.246	0.401	0.955
	linear *fitted*	-0.105	10.170	0.121	0.026	3.189	0.405	0.955
	linear *oracle*	-0.105	9.934	0.123	0.030	3.152	0.419	0.956
	unadjusted *none*	-0.035	9.879	0.083	0.011	3.142	0.429	0.947
small historical sample $(\tilde{n}, n) = (100, 250)$	TMLE *none*	0.055	6.042	0.008	0.033	2.457	0.687	0.950
	TMLE *fitted*	0.009	5.996	−0.017	0.026	2.447	0.677	0.947
	TMLE *oracle*	0.007	4.916	0.018	0.004	2.216	0.739	0.945
	linear *none*	0.120	11.066	0.030	0.026	3.327	0.412	0.954
	linear *fitted*	0.054	10.940	0.001	0.029	3.306	0.415	0.950
	linear *oracle*	0.042	10.812	−0.012	0.028	3.287	0.426	0.952
	unadjusted *none*	0.058	10.415	0.004	0.011	3.226	0.451	0.957
small trial sample $(\tilde{n}, n) = (1000, 100)$	TMLE *none*	0.088	29.818	−0.150	0.208	5.459	0.216	0.944
	TMLE *fitted*	-0.028	14.931	−0.152	0.042	3.862	0.361	0.947
	TMLE *oracle*	0.004	14.034	−0.103	0.041	3.744	0.377	0.949
	linear *none*	0.098	35.371	−0.239	0.273	5.945	0.206	0.931
	linear *fitted*	-0.044	34.146	−0.278	0.300	5.841	0.198	0.929
	linear *oracle*	-0.022	34.290	−0.276	0.296	5.853	0.199	0.934
	unadjusted *none*	0.232	27.199	−0.109	0.061	5.218	0.228	0.945

Appendix D Empirical standard error estimates

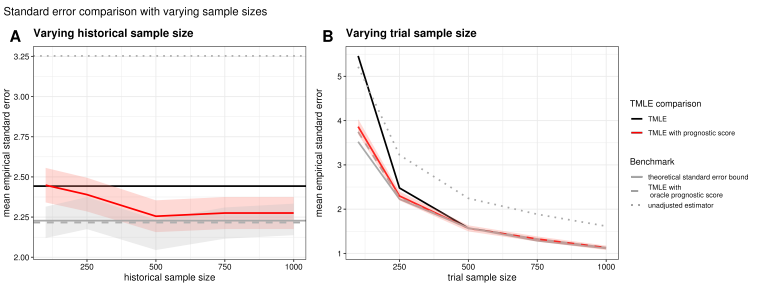

Fig. 1: Mean empirical standard errors across estimators when historical and trial sample sizes are varied using the heterogeneous DGP. When the historical sample size is varied (Appendix Figure 1.A), the trial is fixed at $n = 250$. When the trial size is varied (Appendix Figure 1.B), the historical sample is fixed at $\tilde{n} = 1000$. We show the standard errors with 95% confidence interval for the Monte Carlo simulation as shown in Morris et. al (2019) (1).

Appendix E Case study data summary

Data name	Trial ID	Duration	Titration target (mmol/L)	Blinding type	Number of participants	
					randomized	completed
New RCT	NN9068-4229	26 weeks	4.0-5.0	Open-label	210	206
	NN9068-4228	104 weeks	4.0-5.0	Open-label	504	481
	NN1250-3579	52 weeks	4.0-5.0	Open-label	257	197
	NN1250-3586	26 weeks	4.0-5.0	Open-label	146	136
	NN1250-3672	26 weeks	4.0-5.0	Open-label	230	201
	NN1250-3718	26 weeks	4.0-5.0	Open-label	234	209
	NN1250-3724	26 weeks	4.0-5.0	Open-label	230	206
	NN1250-3587	26 weeks	4.0-5.0	Open-label	278	254
Historical	NN9535-3625	30 weeks	4.0-5.5	Open-label	365	343
	NN2211-1697	26 weeks	$\leq$ 5.0	Double-blinded	34	219
	NN5401-3590	26 weeks	3.9-5.0	Open-label	264	232
	NN5401-3726	26 weeks	3.9-5.0	Open-label	extension of 3590	209
	NN5401-3896	26 weeks	3.9-5.0	Open-label	149	137
	NN1436-4383	26 weeks	4.4-7.2	Double-blinded	122	119
	NN1436-4465	16 weeks	4.4-7.2	Open-label	51	51
	NN1436-4477	78 weeks	4.4-7.2	Open-label	492	477

Tab. 2: Summary of case study data provided by Novo Nordisk A/S. The new RCT data is highlighted in grey. The historical data consists of all the data sets that are not highlighted. The number of participants refers to the number of participants receiving the existing daily insulin treatment IGlar.

Appendix F Summary of continuous measurements of the baseline

Tab. 3: Summary of the continuous baseline covariates.

	Historical sample	New random trial sample
sample size	3311	419
age (years)		
N	3311	419
mean (SD)	57.34 (9.92)	56.67 (10.28)
median	58.00	58.00
min; max	21.00; 85.00	25.00; 83.00
alanine aminotransferase (U/L)		
N	3303	419
mean (SD)	29.51 (17.97)	26.63 (15.48)
median	25.00	23.00
min; max	2.50; 333.00	6.00; 138.00
albumin (g/dL)		
N	3306	419
mean (SD)	4.48 (0.28)	4.51 (0.25)
median	4.50	4.50
min; max	2.50; 5.90	3.80; 5.20
alkaline phosphatase (U/L)		
N	3305	419
mean (SD)	75.99 (23.50)	71.82 (22.31)
median	73.00	68.00
min; max	19.00; 261.00	20.00; 196.00
aspartate aminotransferase (U/L)		
N	3299	419
mean (SD)	23.22 (12.13)	21.38 (9.86)
median	20.00	19.00
min; max	6.00; 227.00	6.00; 89.00
basophils blood (%)		
N	3284	419
mean (SD)	0.53 (0.38)	0.39 (0.22)
median	0.40	0.40
min; max	0.00; 4.40	0.00; 1.60
body mass index (kg/m^2)		
N	3309	419
mean (SD)	30.72 (5.69)	31.22 (4.82)
median	30.22	31.00
min; max	16.01; 56.39	20.00; 43.30
body weight (kg)		
N	3309	419
mean (SD)	86.27 (19.82)	88.30 (17.41)
median	84.90	86.30
min; max	36.30; 171.70	50.98; 145.33
change from baseline to week 26 HbA1c		
N	2642	399
mean (SD)	-1.48 (1.01)	-1.81 (1.01)
median	-1.40	-1.70
min; max	-5.30; 2.60	-6.20 ; 1.20

Continuation of Table 3

	Historical sample	New random trial sample
creatinine (umol/L)		
N	3308	419
mean (SD)	74.14 (18.56)	73.60 (15.64)
median	72.00	72.00
min; max	23.00; 409.00	36.00; 121.00
diabetes duration (years)		
N	3311	419
mean (SD)	9.67 (6.36)	9.55 (6.26)
median	8.67	8.47
min; max	0.30; 49.65	0.44; 34.24
diastolic blood pressure (mmHg)		
N	3309	419
mean (SD)	78.85 (8.53)	79.12 (8.37)
median	80.00	80.00
min; max	47.00; 116.00	57.00; 109.00
eosinophils Blood (%)		
N	3284	419
mean (SD)	2.65 (2.35)	2.56 (2.06)
median	2.10	2.00
min; max	0.00; 43.70	0.00; 15.20
erythrocytes (10^{12}/L)		
N	3297	419
mean (SD)	4.68 (0.45)	5.08 (0.48)
median	4.70	5.00
min; max	3.10; 7.40	3.60; 7.50
fasting plasma glucose (mmol/L)		
N	3268	411
mean (SD)	9.81 (2.65)	9.55 (2.53)
median	9.50	9.20
min; max	2.70; 22.60	3.60; 29.20
haematocrit blood (%)		
N	3264	419
mean (SD)	42.49 (4.16)	45.28 (4.24)
median	42.50	45.50
min; max	22.80; 60.50	31.20; 58.40
hemoglobin A1C at baseline (%)		
N	3311	419
mean (SD)	8.39 (0.92)	8.28 (1.01)
median	8.30	8.10
min; max	6.60; 12.80	6.50; 13.50
high density lipoprotein cholesterol (mmol/L)		
N	3277	410
mean (SD)	1.18 (0.33)	1.21 (0.35)
median	1.14	1.14
min; max	0.21; 3.99	0.31; 2.69
height (m)		
N	3311	419
mean (SD)	1.67 (0.10)	1.68 (0.09)
median	1.67	1.68
min; max	1.36; 2.03	1.43; 2.01
low density lipoprotein cholesterol (mmol/L)		
N	3270	409
mean (SD)	2.46 (0.94)	2.42 (1.01)

Continuation of Table 3

	Historical sample	New random trial sample
median	2.36	2.28
min; max	0.00; 6.73	0.10; 7.10
leukocytes (10^9/L)		
N	3297	419
mean (SD)	7.33 (1.93)	7.93 (2.04)
median	7.10	7.80
min; max	2.80; 17.60	3.60; 15.80
lymphocytes blood (%)		
N	3284	419
mean (SD)	30.00 (7.88)	29.39 (7.66)
median	29.70	28.70
min; max	4.60; 71.00	10.70; 55.10
monocytes blood (%)		
N	3284	419
mean (SD)	5.88 (2.19)	5.83 (2.30)
median	5.70	5.70
min; max	0.00; 21.70	0.50; 17.20
neutrophils blood (%)		
N	3284	419
mean (SD)	60.93 (8.68)	61.83 (9.00)
median	61.10	62.20
min; max	16.60; 91.60	25.20; 86.50
potassium (mmol/L)		
N	3304	419
mean (SD)	4.48 (0.42)	4.53 (0.41)
median	4.49	4.50
min; max	3.10; 7.00	3.30; 6.50
pulse (beats/min)		
N	3310	419
mean (SD)	75.31 (10.00)	75.56 (9.34)
median	75.00	76.00
min; max	45.50; 118.00	52.00; 108.00
sodium (mmol/L)		
N	3303	419
mean (SD)	139.73 (2.81)	140.19 (2.44)
median	140.00	140.00
min; max	121.00; 154.00	132.00; 148.00
systolic blood pressure (mmHg)		
N	3309	419
mean (SD)	131.53 (14.40)	129.69 (13.69)
median	131.00	130.00
min; max	90.00; 200.00	96.00; 171.00
thrombocytes (10^9/L)		
N	3269	419
mean (SD)	240.18 (64.34)	244.27 (64.91)
median	233.00	242.00
min; max	13.00; 611.00	63.00; 477.00
total bilirubin (umol/L)		
N	3304	419
mean (SD)	8.10 (4.33)	8.12 (4.73)
median	7.00	7.00
min; max	0.00; 36.00	1.00; 33.00
total cholesterol (mmol/L)		

Continuation of Table 3

	Historical sample	New random trial sample
N	3284	410
mean (SD)	4.54 (1.13)	4.59 (1.28)
median	4.43	4.43
min; max	0.93; 13.93	2.02; 11.37
triglycerides (mmol/L)		
N	3279	410
mean (SD)	2.07 (1.78)	2.23 (2.36)
median	1.65	1.70
min; max	0.24; 34.25	0.38; 27.80

Appendix G Summary of categorical baseline covariates of the case study

Tab. 4: Summary of the continuous baseline covariates.

	Historical sample		**New random trial sample**	
	N	(%)	N	(%)
sample size	3311		419	
sex				
female	1480	(44.7)	173	(41.3)
male	1831	(55.3)	246	(58.7)
race				
Asian	820	(24.8)	65	(15.5)
Black or African American	183	(5.5)	< 5	
Other	67	(2.0)	< 5	
White	2241	(67.7)	346	(82.6)
smoking status				
current	241	(7.3)	53	(12.6)
never	910	(27.5)	249	(59.4)
previous	378	(11.4)	116	(27.7)
region				
Asia	748	(22.6)	57	(13.6)
Europe	1303	(39.4)	228	(54.4)
North America	1015	(30.7)	89	(21.2)
South Africa	91	(2.7)	0	
South America	154	(4.7)	45	(10.7)
ethnicity				
Hispanic or Latino	461	(13.9)	68	(16.2)
not Hispanic or Latino	2806	(84.7)	351	(83.8)
titration target				
3.9-5.0 mmol/l	410	(12.4)	0	
4.0-5.0 mmol/l	2236	(67.5)	419	(100.0)
4.4-7.2 mmol/l	665	(20.1)	0	
blinding				
double-blinded	122	(3.7)	0	
open-label	3189	(96.3)	419	(100.0)
Biguanides				
yes (continued in trial)	2873	(86.8)	369	(88.1)
yes (discontinued in trial)	248	(7.5)	27	(6.4)
no	190	(5.7)	23	(5.5)
Sulfonylureas				
yes (continued in trial)	436	(13.2)	< 5	
yes (discontinued in trial)	1485	(44.9)	0	
no	1390	(42.0)	> 414	
DPP4				
yes (continued in trial)	278	(8.4)	< 5	
yes (discontinued in trial)	361	(10.9)	> 123	
no	2672	(80.7)	> 285	
other blood glucose lowering drugs				
yes (continued in trial)	< 5		0	

Continuation of Table 4

	Historical sample		New random trial sample	
	N	(%)	N	(%)
yes (discontinued in trial)	> 79		0	
no	> 3220		419	(100.0)
Alpha Glucosidase inhibitor				
yes (continued in trial)	87	(2.6)	0	
yes (discontinued in trial)	69	(2.1)	0	
no	3155	(95.3)	419	(100.0)
combination of blood glucose lowering drug				
yes (continued in trial)	32	(1.0)	0	
yes (discontinued in trial)	36	(1.1)	8	(1.9)
no	3243	(97.9)	411	(98.1)
Thiazolidinediones				
yes (continued in trial)	78	(2.4)	20	(4.8)
yes (discontinued in trial)	29	(0.9)	0	
no	3204	(96.8)	399	(95.2)
SGLT2i				
yes (continued in trial)	175	(5.3)	> 383	
yes (discontinued in trial)	15	(0.5)	> 25	
no	3121	(94.3)	< 5	
GLP-1 receptor agonist				
yes (continued in trial)	79	(2.4)	0	
yes (discontinued in trial)	13	(0.4)	0	
no	3219	(97.2)	419	(100.0)

Appendix H Missing pattern of the case study

To clean and curate the 14 data sets we imputed the HbA1C at week 26 value. For the historical sample the imputation was made using an ANCOVA model with last observed HbA1C measurement before landmark visit, time point of last measurement, baseline HbA1C, discontinuation prior to week 26 indicator and study-id as adjustment covariates. For the new trial data a similar approach was employed. However, in this case the last observed HbA1C measurement before landmark visit week 26, time point of last measurement, baseline HbA1C, discontinuation prior to week 26 indicator, region, treatment indicator and pre-study OADs were used as adjustment covariates. This was done in order to use a similar imputation as used in the original analysis.

After imputing the HbA1C at week 26 value a total 94.4% of the participants had complete data for the combined historical and new trial data. The missingness of the covariates is displayed below. For the covariates we included missingness indicators and respectively imputed covariates using random forest (2). This was done seperately on the historical and new trial data. The normalized root mean square error was 0.218 for continuous covariates and proportion of falsely classified is 0.004 for the historical data sample. The normalized root mean square error was 0.010 for continuous covariates and proportion of falsely classified is 0.023 for the new trial data. The missingness indicators of the historical sample did all overlap with the missingness indicators from NN9068-4229 trial. Since some of the covariates had near zero variance, were colinear or had large absolute correlation with each other we removed some of the covariates.

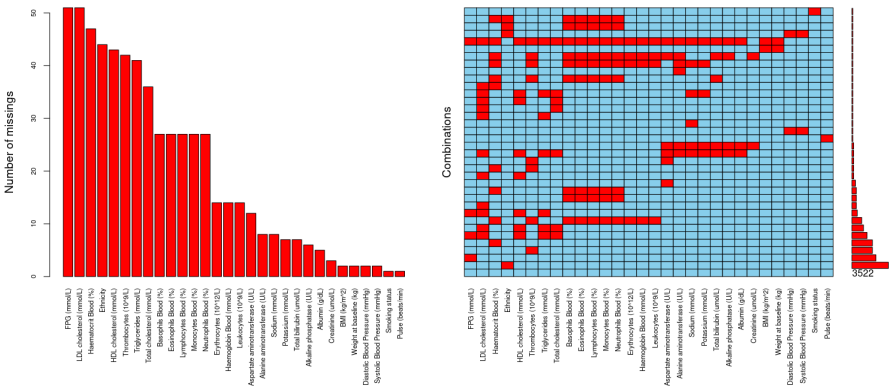

Fig. 2: Number of missing covariates (left) and combination of missingness of covariates (right).

Due to near zero variance, collinearity or high absolute correlation between the covariates, we excluded some of the values. Thus the baseline covariates used in the model where reduced to the following:

- age
- diabetes duration
- body mass index
- HbA1C
- height
- weight
- Alanine aminotransferase
- Albumin
- Alkaline phosphatase
- Aspartate aminotransferase
- Basophils
- Creatinine

- Eosinophils
- Erythrocytes
- fasting plasma glucose
- Haematocrit
- HDL cholesterol
- LDL cholesterol
- Leukocytes
- Lymphocytes
- Monocytes
- Potassium
- Sodium
- Thrombosytes
- Total bilirubin
- Total cholesterol
- Triglycerides
- Diastolic blood pressure
- pulse
- Systolic blood pressure
- country
- sex
- race
- smoking status
- region
- ethnicity
- Biguanides
- DPP4
- SGLT2I
- Previous OADs

Appendix I Case study prognostic score

A linear model was chosen by the super learner for the prognostic score. The explicit form is stated below with all the covariates being found at the baseline visit:
8.064 - 0.005803 * age + 0.006858 * duration of diabetes (years) -0.04832 baseline body mass index -0.6984 baseline HbA1c - 1.597 * baseline height + 0.01584 * baseline weight - 0.0001571 * Alanine aminotransferase -0.1717 * Albumin -0.001101 * Alkaline phosphatase + 0.0007458 * Aspartate aminotransferase + 0.01894 * Basophils + 0.0001417 * Creatinine + 0.00002099 * Eosinophils + 0.07384 * Erythrocytes -0.003852 * Fasting plasma glucose -0.01072 *Haematocrit + 0.1456 * HDL cholesterol -0.02274 * LDL cholesterol + 0.009920 * Leukocytes + 0.002362 * Lymphocytes + 0.001950 * Monocytes + 0.01634 * Potassium + 0.001422 * Sodium + 0.0003064 * Thrombocytes - 0.01402 * Bilirubin -0.009623* Cholesterol + 0.01402 * Triglycerides -0.001623 * Diastolic blood pressure + 0.0003186 * Pulse -0.0003957 * Systolic blood pressure -0.09003 *Canada -0.2896 * Spain -0.1609 * Finland -0.1386 *Hungary + 0.1513 * India + 0.1667 *Russia -0.04964 * Slovakia -0.1002 *Slovenia + 0.01340 * USA + 0.05908 * Male -0.1359 * Race white -0.07616 * current smoker + 0.01871 * never smoked -0.05815 * previous smoker + 0.1683 * Europe -0.1080 * ethnicity not Hispanic or Latino + 0.005146 * Biguanides discontinued + 0.05303 * no Bigunaides used + 0.07725 * Dipeptidyl peptidase-4 inhibitor discontinued.
The covariates are not demeaned. All the categories of the categorical variables included in the model have a main term, meaning that there is no reference group.

References

[1] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in medicine. 2019;38(11):2074-102.
[2] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.