Economics and Human Biology 40 (2021) 100948

Contents lists available at ScienceDirect

Economics and Human Biology

journal homepage: www.elsevier.com/locate/ehb

Regional differences in diabetes across Europe – regression and causal forest analyses

Péter Elek[a,b,c,*], Anikó Bíró[a]

[a] Health and Population "Lendület" Research Group, Centre for Economic and Regional Studies, Budapest, Hungary
[b] Institute of Economics, Corvinus University of Budapest, Hungary
[c] Department of Economics, Eötvös Loránd University (ELTE), Budapest, Hungary

ARTICLE INFO

Article history:
Received 8 July 2020
Received in revised form 4 November 2020
Accepted 10 November 2020
Available online 12 November 2020

JEL classification:
C21
C45
I10
I12
I14

Keywords:
Causal forest
Diabetes
Europe
Health behaviour
SHARE data

ABSTRACT

We examine regional differences in diabetes within Europe, and relate them to variations in socio-economic conditions, comorbidities, health behaviour and diabetes management. We use the SHARE (Survey of Health, Ageing and Retirement in Europe) data of 15 European countries and 28,454 individuals, who participated both in the 4th and 7th (year 2011 and 2017) waves of the survey. First, we estimate multivariate regressions, where the outcome variables are diabetes prevalence, diabetes incidence, and weight loss due to diet as an indicator of management. Second, we study the heterogeneous impact of demographic, socio-economic, health and lifestyle indicators on the regional differences in diabetes incidence with causal random forests.

Compared to Western Europe, the odds of a new diabetes diagnosis over a six-year horizon is 2.2-fold higher in Southern and 2.6-fold higher in Eastern Europe. Adjusting for individual characteristics, the odds ratio decreases to 1.8 in the South-West and to 2.0 in the East-West dimension. These remaining differences are mostly explained by country-specific healthcare indicators. Based on the causal forest approach, the adjusted East-West difference is essentially zero for the lowest risk groups (tertiary education, employment, no hypertension, no overweight) and increases substantially with these risk factors, but the South-West difference is much less heterogeneous. The prevalence of diet-related weight loss around the time of diagnosis also exhibits regional variation. The results suggest that the regional differences in diabetes incidence could be reduced by putting more emphasis on diabetes prevention among high-risk individuals in Eastern and Southern Europe.

1. Introduction

Living with diabetes mellitus is associated with increased all-cause mortality as well as mortality due to cardiovascular disease, chronic lower respiratory diseases, influenza, pneumonia, and kidney disease (Li et al., 2019). More recently, diabetes has been shown to increase the mortality rate and the progression to severe disease in COVID-19 around twofold (Huang et al., 2020).

In this paper, our aim is to document regional differences in the prevalence and incidence of diabetes across Europe, and to relate these differences to variations in socio-economic conditions, comorbidities, health behaviour and diabetes management. Around 8.9% of Europeans aged 20-79 years live with diabetes, and 8.5% of all deaths is attributable to diabetes and its complications (IDF, 2019). However, the regional distribution is very uneven: prevalence shows a more than twofold and mortality a more than fourfold difference even across the member states of the European Union (Whiting et al., 2011; IDF, 2019). It is well known that genetics, lifestyle, diet and the healthcare system all influence the incidence and mortality in diabetes, and these risk factors are unevenly distributed across the population of European countries (Tamayo et al., 2014). In particular, the roles of socio-economic inequalities (Agardh et al., 2011; Espelt et al., 2013), of the body mass index (e.g. Narayan et al., 2007) and of lifestyle changes (e.g. Diabetes Prevention Program Research Group, 2002) are well documented. However, less is known about the relative role of these factors in explaining the variation of diabetes across Europe. Our study aims to fill this gap.

The risk factors for diabetes are highly correlated and may influence diabetes prevalence and incidence in a nonlinear and non-additive way. For instance, the combination of obesity, hypertension, slightly elevated blood sugar and abnormal cholesterol level markedly increases the risk of cardiovascular disease and the transition rate to overt diabetes. This is the rationale

* Corresponding author. 1097 Budapest, Tóth Kálmán utca 4., Hungary.
E-mail addresses: elek.peter@krtk.hu (P. Elek), biro.aniko@krtk.hu (A. Bíró).

http://dx.doi.org/10.1016/j.ehb.2020.100948

behind the diagnosis of the metabolic syndrome, which is defined, roughly, when a patient has at least three risk factors out of the above four. Some studies argue that metabolic syndrome is more than its parts in terms of cardiovascular or overt diabetes risk (but see e.g. Kassi et al., 2011 for a review of controversies), suggesting interaction effects between the risk factors. We train a causal random forest developed by Wager and Athey (2018) and Athey et al. (2019) to investigate heterogeneity in the adjusted regional differences in diabetes incidence. Specifically, we analyse how the regional differences in diabetes incidence vary by demographic, socio-economic, health and lifestyle indicators. Compared to a traditional full interaction linear regression model, the main benefit of the causal forest method is the gain in statistical power due to the automated choice of heterogeneities to be included in the model. Finally, we investigate regional differences in the change in health behaviour (the probability of weight loss due to diet) around the time of diabetes diagnosis. The results shed light on the origins of the marked cross-country differences in diabetes throughout Europe.

We use the Survey of Health, Ageing and Retirement in Europe (SHARE) (Börsch-Supan, 2019). SHARE is a cross-national European panel database of micro data on demographic, socio-economic, labour market, health and lifestyle information of individuals aged 50 or older, hence it is a convenient database for analysing all important diabetes-related factors simultaneously. Indeed, a number of studies have used SHARE for diabetes research. Based on SHARE data, Rodriguez-Sanchez and Cantarero-Prieto (2019) show a positive association of diabetes with hospital admissions and death, while Espelt et al. (2013) find that education is inversely associated with diabetes prevalence and (for women) with diabetes incidence. Diabetes is known to increase the rate of labour force exit by around 30% (Rumball-Smith et al., 2014) and the probability of disability benefits more than twofold (Kouwenhoven-Pasmooij et al., 2016). Bashkin et al. (2018) also use SHARE data to show that the positive association between diabetes and depression is no longer significant after adjusting for a rich set of individual characteristics.

We make several contributions to the existing literature. First, beyond examining diabetes prevalence, we also analyse the transition to diabetes over a six-year horizon, a sufficiently long time period to measure the effect of the explanatory variables. Second, we investigate how the prevalence and incidence differences between the three regions of Europe vary by individual risk factors and apply the novel causal forest methodology to answer this question. Finally, we relate weight loss – an indicator of diabetes management – to the patterns of transition to diabetes.

2. Data

The SHARE surveys were conducted in seven waves, starting in 2004, and the currently last wave was taken in 2017.[1] The number of participating countries gradually expanded from 12 to 27 to include new EU member states as well. We exploit the panel nature of the survey by using waves 4 and 7, which were taken six years apart (2011 and 2017), hence transition to diabetes can be reliably examined on them. We split the 15 countries that appear in both waves into three groups: West [including North] (Austria, Belgium, Denmark, France, Germany, Sweden, Switzerland); South (Italy, Portugal, Spain); East (Czech Republic, Estonia, Hungary, Poland, Slovenia).[2] We use calibrated weights to avoid bias due to unit nonresponse and panel attrition (see Malter and Börsch-Supan, 2015 for details).

In our analysis, we treat a person as having diabetes if he / she answered "yes" to any of the following two questions: (1) "Has a doctor ever told you that you had / Do you currently have diabetes or high blood sugar?" (2) "Do you currently take drugs at least once a week for diabetes or high blood sugar?" We examine diabetes prevalence, i.e. the binary indicator of having diabetes in wave 7; and diabetes incidence (transition to diabetes), i.e. the binary indicator of having diabetes in wave 7 among those who did not have diabetes in wave 4. We do not distinguish between Type 1 and Type 2 diabetes, but around 90% of the prevalence and the overwhelming majority of incidence above 50 years belongs to the latter category (IDF, 2019).

Other variables – which we use as explanatory variables – include region, demographic and socio-economic characteristics (gender, age, years of education, employment status), body mass index (BMI, calculated from self-reported height and weight, and then categorised into normal weight ($BMI < 25$), overweight ($25 \leq BMI < 30$), obesity ($30 \leq BMI$)[3] and, as a subgroup, severe obesity ($35 \leq BMI$)), comorbidities (hypertension and high cholesterol, measured by drug use on these conditions, and having ever been diagnosed with heart attack or stroke) and lifestyle factors (binary indicators of smoking now; playing sports at least once a week; eating fruits or vegetables daily). We use the explanatory variables from wave 4. The dataset also contains the self-reported binary indicator of having lost weight due to diet during the past 12 months in wave 7.

We merge three country-specific healthcare indicators to the SHARE data: total healthcare spending per GDP (source: Eurostat, 2020); the number of physicians per 1,000 inhabitants (source: WHO, 2020); and the share of the population aged 16 and above who report unmet needs for medical care due to financial reasons, waiting lists or having to travel too far (source: Eurostat, 2020). The indicators refer to year 2011 (the time of wave 4), except for health spending per GDP, which refers to year 2013 (due to missing data in 2011). Our aim with these indicators is to capture healthcare availability and quality. While the number of physicians and the prevalence of unmet needs are direct measures of healthcare availability, the indicator of healthcare spending can serve as a proxy for healthcare quality due to its known relation to advancements in medical technology (Beilfuss and Thornton 2016; Newhouse, 1992) and lower avoidable mortality (Heijink et al., 2013).

[1] This paper uses data from SHARE Waves 1, 2, 3, 4, 5, 6 and 7 (DOIs: 10.6103/SHARE.w1.700, 10.6103/SHARE.w2.700, 10.6103/SHARE.w3.700, 10.6103/SHARE.w4.700, 10.6103/SHARE.w5.700, 10.6103/SHARE.w6.700, 10.6103/SHARE.w7.700), see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been primarily funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812) and FP7 (SHARE-PREP: No211909, SHARE-LEAP: No227822, SHARE M4: No261982). Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

[2] We use data from waves 4 and 7 to ensure that at least three countries are present from each region and a sufficient number of transitions to diabetes is observed. Hungary, Poland and Portugal were not included in wave 5, Hungary did not appear in wave 6, either. Using data from waves 4 and 6, waves 5 and 7 or waves 6 and 7 would reduce the number of observed transitions by 15%, 17% and 34%, respectively. The variation in country coverage and survey content hinders us from conducting a panel analysis with more than two waves.

[3] Only 1% of the respondents in wave 4 are underweight ($BMI < 18.5$), whom we include in the normal weight category.

3. Methods

3.1. Multivariate regressions

We fit linear probability and logit models on diabetes prevalence (i.e. being diabetic in wave 7) and incidence (i.e. becoming diabetic between waves 4 and 7), respectively. We include gradually more control variables beyond the regional dummies (or in some specifications the country dummies) to examine their confounding effect on regional / cross-country differences. First, we add the demographic and socio-economic indicators (age, age squared, gender, education categories and employment). Then, we extend the models with indicators of health status (BMI categories, hypertension, high blood cholesterol) and health behaviour (smoking, weekly sports activity and daily fruit or vegetable consumption).[4] To reduce the problem of reverse causality on the individual level, we use the explanatory variables from wave 4. As the last extension, we add the country-specific healthcare indicators to the explanatory variables to check whether they explain the remaining part of regional differences.

We also investigate the change in health behaviour around the time of diabetes diagnosis. Specifically, we estimate linear and logit models of the probability of weight loss due to diet in wave 7. We fit these models on the sample of individuals who were not diabetic in wave 4, and use the interaction of the regional dummies with the wave 7 diabetes dummy (and other controls) to investigate regional heterogeneities in the prevalence of weight loss due to the diagnosis of diabetes. As a supplementary analysis, to understand the co-movement of diabetes diagnosis and weight loss due to diet, we add waves 5 and 6 of the SHARE data to our sample, look at the two-year transitions to diabetes for the available countries and analyse the prevalence of weight loss due to diet concurrently, as well as one or two waves before and one or two waves after the diagnosis (i.e. up to four years before and four years after it).

3.2. Causal forests

In order to analyse the heterogeneous effect of the risk factors on the regional differences in diabetes incidence, we train causal forests separately on East-West and South-West differences (in each analysis we omit the third category from the sample). Specifically, let W_i be the regional dummy (which takes one for East or South and zero for West) and X_i denote the individual-level demographic, socio-economic, health- and lifestyle-related control variables. Let $Y_i(1) = Y_i(W_i = 1)$ and $Y_i(0) = Y_i(W_i = 0)$ be the potential outcomes, i.e. the diabetes status of a particular person in the (imagined) situation that she is in Eastern (Southern) or in Western Europe, respectively. (We set Western Europe as the reference category because diabetes incidence is the lowest there.) We seek to estimate the conditional "treatment" effect

$$\tau(x) = E(Y_i(1) - Y_i(0)|X_i = x)$$

assuming unconfoundedness and overlap. The unconfoundedness assumption states that $\{Y_i(1), Y_i(0)\}$ are independent from W_i conditional on the value of X_i, while overlap means that $0 < p(x) < 1$, where $p(x) = \Pr(W_i = 1|X_i = x)$ is the propensity score.

The fundamental problem of estimating treatment effects lies in the fact that for each i we only observe either $Y_i(1)$ or $Y_i(0)$, not both. Still, unconfoundedness ensures that we can treat observations with similar x values as if they came from a randomized experiment, hence can estimate $\tau(x)$ by comparing realized $Y_i(1)$ and $Y_i(0)$ outcomes for similar $X_i = x$ values. Under the overlap assumption, such realised outcomes are available for both groups.

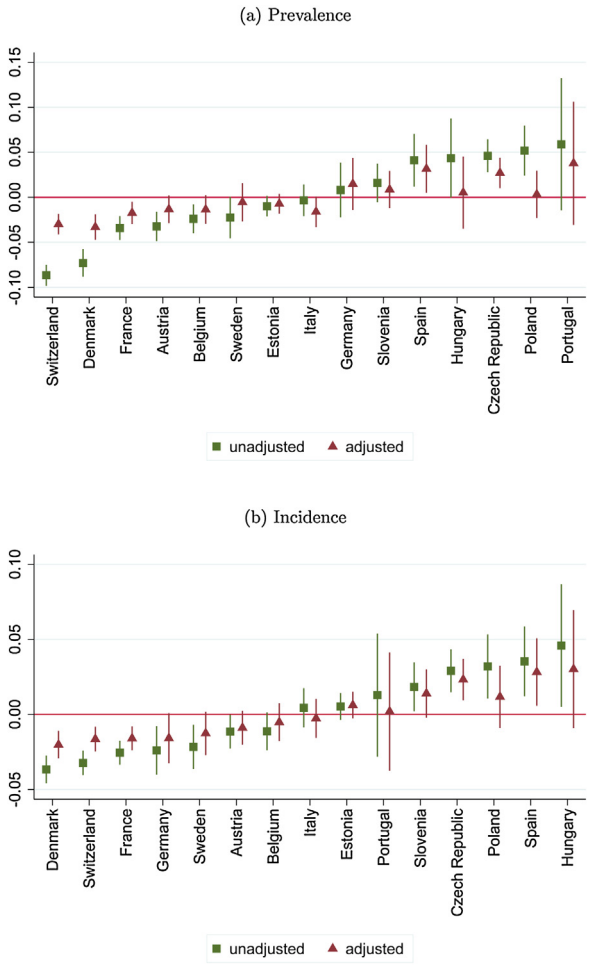

Fig. 1. Unadjusted and adjusted diabetes prevalence and incidence by countries, deviations from the mean. Adjustment is made by controlling for the individual-specific variables listed in Tables 2 and 3 .

Since in our setting the regional dummies do not have a clear causal interpretation, we interpret $\tau(x)$ in this paper merely as the heterogenous regional "effect" after controlling for the observable variables. Hence, instead of the usual "treatment effect" we can use the term "predictive effect" (following e.g. Chernozhukov et al., 2018a,b) or "adjusted difference" (following the epidemiological literature).

Traditional regression methods estimate the effect of W_i by adjusting for X_i in a parametric way, while nearest neighbour methods explicitly search for observations with similar X_i but different W_i values. Heterogeneous effects (varying $\tau(x)$) can be estimated in the regression setting by including interaction terms between W_i and X_i, or, similarly, by fitting separate regression models to different subsamples defined according to the values of X_i. However, if the number of potential interaction terms is substantial, model selection (i.e., deciding which interaction terms should be used) can be difficult because of multicollinearities.

The causal forest, developed by Wager and Athey (2018) and Athey et al. (2019), is a promising new and automated way to estimate heterogeneous treatment (or predictive) effects. According to the simulations of Wager and Athey (2018), it provides much

[4] We experimented with the addition of a rich set of further controls such as household size, marital status, childhood health or alcohol consumption. These controls turned out to be statistically insignificant and their inclusion in the models did not change the coefficients of the regional dummies.

Table 1
Descriptive statistics of the variables used in Table 2.

	West		South		East		South-West diff	East-West diff
	mean	SE(mean)	mean	SE(mean)	mean	SE(mean)	p-value	p-value
diabetes (wave 7)	0.127	0.003	0.183	0.006	0.201	0.004	0.000	0.000
age	63.461	0.087	62.590	0.161	63.001	0.086	0.005	0.080
female	1.559	0.004	1.551	0.008	1.595	0.005	0.604	0.008
years of education 0-8	0.221	0.004	0.630	0.008	0.288	0.004	0.000	0.000
years of education 9-12	0.395	0.004	0.165	0.006	0.497	0.005	0.000	0.000
years of education 13+	0.383	0.004	0.205	0.006	0.215	0.004	0.000	0.000
employment	0.363	0.004	0.312	0.007	0.269	0.004	0.001	0.000
BMI -25	0.426	0.004	0.338	0.007	0.272	0.004	0.000	0.000
BMI 25-30	0.397	0.004	0.456	0.008	0.413	0.005	0.000	0.234
BMI 30-35	0.137	0.003	0.157	0.006	0.228	0.004	0.027	0.000
BMI 35+	0.041	0.002	0.048	0.003	0.086	0.003	0.176	0.000
hypertension	0.353	0.004	0.367	0.008	0.506	0.005	0.282	0.000
high cholesterol	0.221	0.004	0.233	0.007	0.219	0.004	0.313	0.875
ever had heart attack	0.093	0.003	0.082	0.004	0.153	0.003	0.122	0.000
ever had stroke	0.026	0.001	0.022	0.002	0.043	0.002	0.292	0.001
smoker	0.179	0.003	0.192	0.006	0.235	0.004	0.296	0.000
sports weekly	0.520	0.004	0.367	0.008	0.425	0.005	0.000	0.000
fruit or vegetable daily	0.812	0.003	0.828	0.006	0.623	0.005	0.178	0.000
country-specific indicators								
health spending per GDP (%)	11.113	0.004	9.000	0.001	7.139	0.006	0.000	0.000
physicians per 1,000 population	3.469	0.005	3.880	0.001	2.914	0.005	0.000	0.000
ratio with unmet healthcare needs (%)	1.776	0.006	2.895	0.040	3.891	0.027	0.000	0.000
number of individuals	13,253		4,086		11,115			

Apart from wave 7 diabetes, all indicators are measured in wave 4.
The last two columns show the results of t-test of equality across country groups.

better mean-squared error than e.g. classical nearest neighbour methods. (In Appendix B we compare the heterogeneities found by the causal forest method to those obtained from subsample-specific regression estimates.)

The causal forest builds on the random forest algorithm, which was designed for pure forecasting purposes, i.e. for estimating conditional expectations $m(x) = E(Y_i|X_i = x)$. Forecasts from random forests are obtained by averaging forecasts from many individual decision trees, each of which is fitted on a bootstrapped subsample of the original sample (called bootstrap aggregation or bagging), with one additional twist: during each split of a tree the partitioning variable may only be chosen from a random subset of the full variable list. A split of a tree is carried out by maximising the heterogeneity of the predictions across the resulting two child nodes. (For more details on random forests see e.g. Hastie et al. (2009)).

Instead of estimating $m(x)$, causal forests focus on the estimation of $\tau(x)$. The basic idea is that the conditional treatment (or predictive) effect of W_i at x can be estimated by taking the difference of the average outcomes of observations with $W_i = 1$ and $W_i = 0$ within the leaf L of the tree that contains x:

$$\hat{\tau}(x) = \overline{Y}_{\{W_i=1;X_i\in L\}} - \overline{Y}_{\{W_i=0;X_i\in L\}}. \quad (1)$$

Appendix A shows the details of the causal forest methodology, which we implement with the R package *grf* (Tibshirani et al., 2019).[5] We use the automatic tuning procedure of the package to determine the parameters of the forest (e.g. the minimum leaf size), apart from the number of trees grown, which is set as 32,000 for the East-West comparison and 64,000 for the South-West comparison (the latter being larger due to the smaller sample size).[6] After growing the forest, we estimate average treatment (or predictive) effects on subsamples split according to the presence or absence of various risk factors. Following Athey and Wager (2019) and Farrell (2015), here we use an augmented inverse propensity weighting (AIPW) correction.

Finally, we evaluate the fit of the estimated causal forests in three ways. First, we check the overlap assumption by looking at whether the propensity scores are bounded away from zero and one. Second, we investigate covariate balance by comparing the inverse-propensity weighted averages of the explanatory variables across the two groups. Third, we implement the "best linear predictor" method of Chernozhukov et al. (2018a,b). (See Appendix A for details.)

4. Results

4.1. Prevalence

Fig. 1a and the first column of Table 2 show unadjusted (raw) differences in diabetes prevalence across countries and regions, respectively, referring to the population aged at least 50 years as sampled in SHARE. Prevalence is significantly higher than average in Poland, the Czech Republic and Spain, and lower in Switzerland, Denmark, Austria, France, Belgium and Sweden. Taken the countries together, the prevalence exceeds the Western European average (12.7%) by 7.4% points in Eastern and by 5.6% points in Southern Europe.

Descriptive statistics in Table 1 show that, compared to the Western European population, Eastern Europeans on average have less education, are less likely to be employed, have a higher BMI (particularly in the obese and severely obese range), are more often diagnosed with hypertension, are more likely to have ever had a heart attack or stroke, smoke more often; but play sports at least weekly or eat fruit or vegetable daily in a smaller proportion. Southern Europeans have less education, are less likely to be employed, have a higher BMI (in the overweight range) and less often play sports than Western Europeans, but otherwise the differences are smaller than in the East-West dimension.

[5] See also the technical reference of the package (https://github.com/grf-labs/grf/blob/master/REFERENCE.md), section 6.2. of Athey et al. (2019) or section 1.3. of Athey and Wager (2019).

[6] We tried various other specifications for the causal forest such as using the baseline built-in parameters or tuning only a subset of the parameters, but the conclusions did not change.

Table 2
OLS and logit models of diabetes prevalence in SHARE wave 7.

Dep. var.: prevalence	linear probability model effects					logit model odds ratios		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
South	0.056***	0.038***	0.030***	0.023[a]	1.541***	1.363***	1.334***	1.198
	[0.010]	[0.009]	[0.009]	[0.014]	[0.113]	[0.097]	[0.101]	[0.141]
East	0.074***	0.061***	0.024**	0.003	1.736***	1.587***	1.259***	0.989
	[0.011]	[0.011]	[0.010]	[0.025]	[0.127]	[0.119]	[0.099]	[0.207]
(age-50)/10		0.025	0.001	0.010		1.248	1.023	1.100
		[0.018]	[0.017]	[0.022]		[0.193]	[0.159]	[0.219]
(age-50)/10 squared		−0.005	−0.000	−0.002		0.955	1.000	0.987
		[0.005]	[0.004]	[0.004]		[0.037]	[0.039]	[0.038]
female		−0.053***	−0.046***	−0.045***		0.659***	0.654***	0.660***
		[0.009]	[0.008]	[0.008]		[0.043]	[0.045]	[0.045]
education 9-12 years		−0.022**	−0.015	−0.016		0.871[a]	0.916	0.909
		[0.011]	[0.010]	[0.010]		[0.067]	[0.073]	[0.073]
education 13+ years		−0.051***	−0.020**	−0.021**		0.662***	0.838[a]	0.830**
		[0.010]	[0.010]	[0.010]		[0.056]	[0.076]	[0.075]
employment		−0.096***	−0.069***	−0.069***		0.408***	0.473***	0.476***
		[0.014]	[0.013]	[0.013]		[0.050]	[0.058]	[0.058]
BMI 25-30			0.053***	0.052***			1.921***	1.905***
			[0.008]	[0.008]			[0.167]	[0.165]
BMI 30-35			0.159***	0.156***			3.735***	3.669***
			[0.014]	[0.014]			[0.367]	[0.359]
BMI at or above 35			0.286***	0.281***			6.746***	6.544***
			[0.025]	[0.025]			[0.889]	[0.859]
hypertension			0.061***	0.062***			1.611***	1.625***
			[0.009]	[0.009]			[0.114]	[0.114]
high cholesterol			0.079***	0.078***			1.716***	1.706***
			[0.011]	[0.012]			[0.126]	[0.127]
ever had heart attack			0.027	0.026			1.123	1.119
			[0.017]	[0.016]			[0.110]	[0.109]
ever had stroke			0.020	0.020			1.107	1.102
			[0.025]	[0.025]			[0.165]	[0.164]
smoker			0.010	0.009			1.103	1.094
			[0.011]	[0.011]			[0.106]	[0.105]
sports weekly			−0.033***	−0.034***			0.746***	0.741***
			[0.008]	[0.008]			[0.053]	[0.053]
fruit or veg. daily			−0.005	−0.004			0.949	0.953
			[0.010]	[0.010]			[0.079]	[0.080]
health spending / GDP (%)				−0.009				0.914
				[0.006]				[0.052]
physicians / 1,000 pop.				−0.006				0.957
				[0.007]				[0.058]
ratio with unmet needs (%)				−0.008***				0.941***
				[0.002]				[0.017]
constant	0.127***	0.251***	0.159***	0.281***	0.145***	0.359***	0.151***	0.479
	[0.005]	[0.028]	[0.027]	[0.095]	[0.006]	[0.075]	[0.034]	[0.379]

Number of observations: 28,454. All explanatory variables are measured in wave 4.
Standard errors in brackets (OLS: robust),
*** p<0.01.
** p<0.05.
[a] p<0.1.

Looking at the country-specific indicators, health spending per GDP and the density of physicians are lower, while the prevalence of unmet needs is higher in the East than in the West. In the South, these indicators are in between, apart from the number of physicians, which is the highest there.

According to Table 2, the unadjusted East-West difference of 7.4 %points (odds ratio [OR] = 1.74) is only slightly reduced by controlling for demographic and socio-economic variables (age, gender, years of education and employment) but decreases to less than half (to 2.4 %points, OR=1.26) by controlling further for health-related (BMI, hypertension, high cholesterol, heart attack, stroke) and lifestyle factors. The South-West difference decreases by more than one-third, from 5.6% points to 3.0 %points (OR from 1.54 to 1.33) after controlling for these variables. According to Fig. 1a, the unadjusted and adjusted differences are far from each other in Switzerland and Denmark, where a substantial portion of the better than average prevalence is explained by the more favourable distribution of the risk factors. On the other hand, in Hungary and Poland, the bulk of the worse than average prevalence is explained by the worse values of the observed risk factors.

As columns (4) and (8) in Table 2 indicate, once we add the three country-specific indicators to the regression (health spending per GDP, physicians per capita, prevalence of unmet needs), the East-West difference in diabetes prevalence disappears and the South-West difference decreases substantially. Thus, differences in healthcare availability largely explain the residual differences in prevalence.

4.2. Incidence

In the following, we focus on incidence (i.e. on the transition to diabetes between waves 4 and 7) because risk factors measured before the diagnosis (in wave 4) are more plausibly exogenous. Fig. 1b and the first column of Table 3 show how the transition rate from non-diabetes to diabetes varies across countries and regions. Incidence is significantly higher than average in Hungary, Spain,

Table 3
OLS and logit models of new diabetes diagnosis between SHARE waves 4 and 7.

Dep. var.: incidence	linear probability model effects					logit model odds ratios		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
South	0.043***	0.035***	0.032***	0.022**	2.195***	1.953***	1.847***	1.391[a]
	[0.007]	[0.007]	[0.007]	[0.010]	[0.245]	[0.215]	[0.210]	[0.248]
East	0.059***	0.055***	0.039***	0.010	2.642***	2.513***	2.009***	1.036
	[0.009]	[0.009]	[0.009]	[0.020]	[0.319]	[0.303]	[0.244]	[0.346]
(age-50)/10		−0.012	−0.018	−0.023		0.819	0.716	0.655
		[0.014]	[0.014]	[0.017]		[0.187]	[0.167]	[0.192]
(age-50)/10 squared		0.005	0.007[a]	0.006[a]		1.080	1.123**	1.114[a]
		[0.003]	[0.003]	[0.003]		[0.059]	[0.064]	[0.062]
female		−0.032***	−0.027***	−0.027***		0.574***	0.601***	0.605***
		[0.006]	[0.006]	[0.006]		[0.059]	[0.063]	[0.063]
education 9-12 years		−0.012	−0.009	−0.009		0.855	0.891	0.886
		[0.008]	[0.008]	[0.008]		[0.107]	[0.114]	[0.113]
education 13+ years		−0.020**	−0.008	−0.009		0.713**	0.858	0.857
		[0.008]	[0.008]	[0.008]		[0.101]	[0.126]	[0.126]
employment		−0.034***	−0.027**	−0.027**		0.507***	0.541***	0.542***
		[0.011]	[0.011]	[0.011]		[0.101]	[0.109]	[0.108]
BMI 25-30			0.026***	0.026***			2.068***	2.038***
			[0.005]	[0.005]			[0.269]	[0.266]
BMI 30-35			0.082***	0.081***			4.005***	3.948***
			[0.012]	[0.012]			[0.606]	[0.595]
BMI at or above 35			0.118***	0.116***			5.462***	5.306***
			[0.023]	[0.023]			[1.135]	[1.105]
hypertension			0.015**	0.016**			1.301**	1.326***
			[0.007]	[0.007]			[0.138]	[0.140]
high cholesterol			0.017[a]	0.015[a]			1.300**	1.261[a]
			[0.009]	[0.009]			[0.156]	[0.154]
ever had heart attack			−0.009	−0.009			0.878	0.879
			[0.010]	[0.010]			[0.117]	[0.117]
ever had stroke			0.037[a]	0.036			1.473[a]	1.454[a]
			[0.022]	[0.022]			[0.323]	[0.319]
smoker			0.009	0.009			1.182	1.178
			[0.008]	[0.008]			[0.176]	[0.176]
sports weekly			−0.007	−0.007			0.860	0.860
			[0.006]	[0.006]			[0.095]	[0.096]
fruit or veg. daily			−0.009	−0.008			0.857	0.865
			[0.008]	[0.008]			[0.104]	[0.106]
health spending / GDP (%)				−0.009[a]				0.822**
				[0.005]				[0.074]
physicians / 1,000 pop.				−0.009				0.895
				[0.005]				[0.083]
ratio with unmet needs (%)				−0.005***				0.923***
				[0.002]				[0.024]
constant	0.040***	0.117***	0.082***	0.233***	0.041***	0.149***	0.070***	1.237
	[0.003]	[0.022]	[0.022]	[0.076]	[0.003]	[0.049]	[0.025]	[1.549]

Number of observations: 24,967. All explanatory variables are measured in wave 4.
Standard errors in brackets (OLS: robust),
*** p<0.01.
** p<0.05.
[a] p<0.1.

Poland, the Czech Republic, and lower in Denmark, Switzerland, France, Germany, Sweden and Austria. Six-year incidence is by 5.9 %points (OR = 2.64) higher in Eastern and by 4.3 %points (OR = 2.20) higher in Southern Europe than in Western Europe (4.0%). According to Table 3, controlling for the individual-level variables measured at wave 4 reduces the East-West difference to 3.9 % points (OR = 2.01) and the South-West difference to 3.2 %points (OR = 1.85).

Among the control variables, the three additional health-related components of the metabolic syndrome all increase the rate of transition to overt diabetes. Even overweight ($25 \leq BMI < 30$), which characterises more than 40% of the 50+ population, is a significant risk factor (OR = 2.1), while the two classes of obesity have a markedly larger effect (OR = 4.0 and 5.5, not significantly different from each other). Previous hypertension and high blood cholesterol have ORs around 1.3. Female gender and employment are associated with strongly reduced diabetes incidence, while measured lifestyle factors have only marginally significant effects.

Just as in the case of diabetes prevalence, the bulk of the remaining East-West difference and a large portion of the South-West difference in diabetes incidence is explained by the three country-specific healthcare indicators (columns (4) and (8) of Table 3).

4.3. Heterogeneity in incidence

The descriptive plots of Fig. 2 show that hypertension, high blood cholesterol and high BMI are associated with a higher probability of new diabetes diagnosis in all three regions, but to varying degrees. For instance, the association of hypertension and high cholesterol with the diagnosis seems to be more pronounced in Eastern Europe than elsewhere.

To analyse the heterogeneity of regional effects, we train causal forests as explained in section 3.2, using the same individual-level control variables as in column (3) of Table 3. The causal forests yield very similar estimates for the average adjusted regional differences in diabetes incidence as the controlled linear

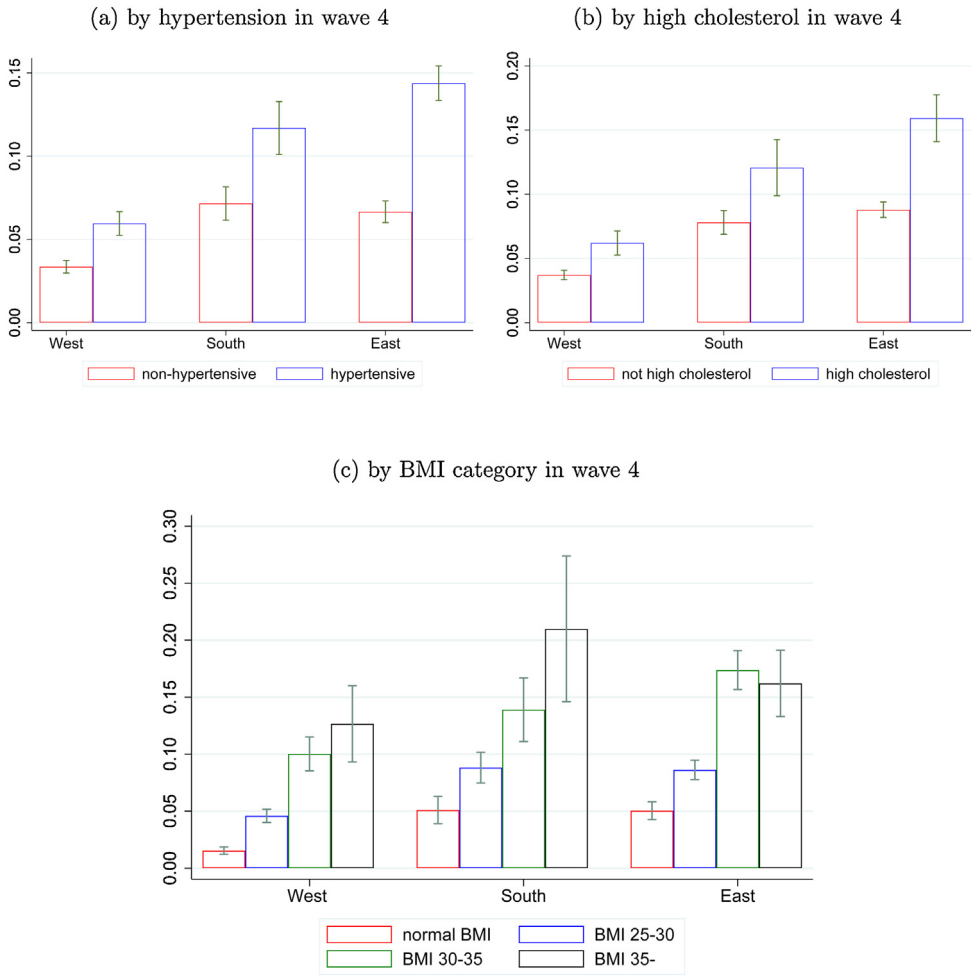

Fig. 2. Transition probabilities to diabetes between SHARE waves 4 and 7, by hypertension, high cholesterol and BMI category measured in wave 4.

probability model in Table 3: 3.8 %points (S.E. = 0.6 %point) for the East-West difference and 3.9 %points (S.E.= 0.4 %point) for the South-West difference. However, the value added of the causal forest approach is that it automatically yields effect estimates for each individual, so that they can be aggregated by different risk factors.

The subsample-specific adjusted regional differences (calculated from equation (4) in Appendix A), their 95% confidence intervals and the statistical significance of the between-group variations are displayed in Fig. 3.[7] The adjusted East-West difference is significantly higher for individuals with lower education level, without employment in wave 4, with higher BMI, with previous hypertension or high cholesterol, in such a way that the least vulnerable groups have essentially no excess transition risk to diabetes in the East compared to the West. For instance, the effect estimate is 1.1 %point (not significantly different from zero at the 10% level) for those with more than 12 years of schooling but 5.5 % points for those with at most 8 years, or 1.3 %points (not significantly different from zero at the 5% level) for those without hypertension and 7.0 %points for those with it. Among lifestyle factors, smoking increases the excess risk significantly at the 5% level. Meanwhile, effect heterogeneity is not significant across gender, age, previous heart attack or weekly sports activity.

Compared to the the East-West dimension, heterogeneity is much less pronounced in the South-West dimension (Fig. 3b), where the effect estimate is statistically significantly positive in each subsample. Significant heterogeneity is observed only by level of education, previous employment status, BMI and sports activity.

For robustness check, Appendix B displays subsample-specific regional effects that were estimated by OLS regressions run on different subsamples, using the same control variables as in column (3) of Table 3. (For illustration, the subsamples were defined by health status.) The OLS point estimates are similar to the causal forest ones but have larger standard errors because the causal forest methodology automatically chooses the heterogeneities that should be included in the model (while the subsample-specific OLS models implicitly use many interaction terms between the control variables).

Appendix C contains the goodness-of-fit analysis of the estimated causal forests. According to Fig. C1, the propensity scores are mainly between 0.05 and 0.95, hence the overlap assumption holds both for the East-West and the South-West model. Table C1 shows that the large (standardized) differences in the explanatory variables (especially in BMI, hypertension and

[7] The figures do not show heterogeneities by the presence or absence of previous stroke because the ratio of individuals who ever had a stroke is only 2-4% hence the regional effects for them are very imprecisely estimated.

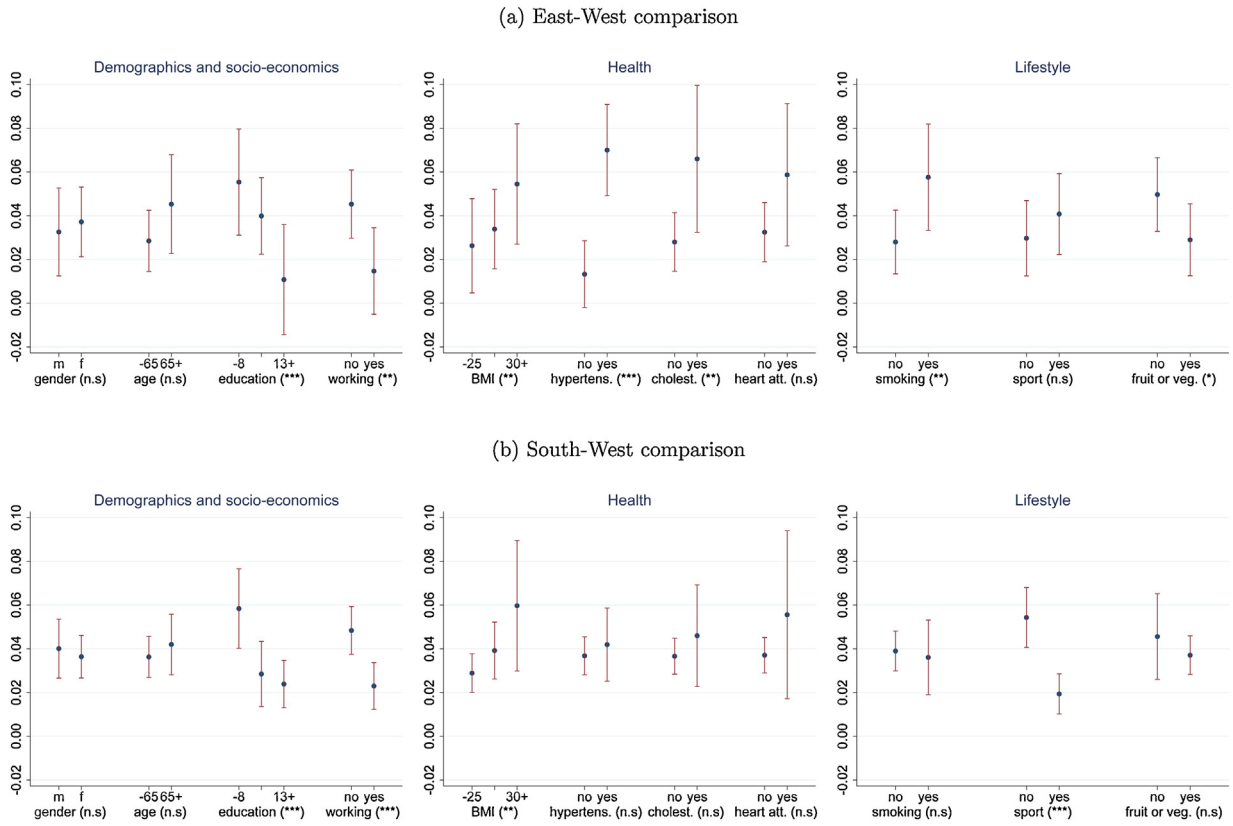

Fig. 3. Average individual-level adjusted regional differences by various risk factors, with 95% confidence intervals and with the statistical significance of the between-group variation (*** p<0.01, ** p<0.05, * p<0.1, n.s p≥0.1).

Table 4
Models of the probability of reporting weight loss due to diet in SHARE wave 7.

	linear probability model effects		logit model odds ratios	
	(1)	(2)	(3)	(4)
South	−0.007	−0.025***	0.894	0.650**
	[0.012]	[0.010]	[0.175]	[0.112]
East	−0.051***	−0.054***	0.277***	0.268***
	[0.007]	[0.007]	[0.046]	[0.046]
West × diabetes in wave 7	0.077**	0.101**	2.245***	2.826***
	[0.035]	[0.040]	[0.644]	[0.849]
South × diabetes in wave 7	0.007	0.022	1.109	1.463
	[0.025]	[0.026]	[0.429]	[0.641]
East × diabetes in wave 7	0.038[a]	0.035[a]	2.886***	2.717**
	[0.019]	[0.021]	[1.069]	[1.098]
(age-50)/10		0.005		1.274
		[0.021]		[0.503]
(age-50)/10 squared		−0.005		0.880
		[0.004]		[0.081]
female		0.021**		1.456**
		[0.009]		[0.217]
education 9-12 years		−0.012		0.800
		[0.010]		[0.149]
education 13+ years		0.006		1.101
		[0.010]		[0.186]
employment		−0.009		0.861
		[0.011]		[0.153]
constant	0.072***	0.062**	0.078***	0.056***
	[0.006]	[0.028]	[0.007]	[0.027]

Sample: non-diabetic and overweight or obese in wave 4. Number of observations: 10,406.
The explanatory variables are measured in wave 4, apart from diabetes in wave 7.
Standard errors in brackets (OLS: robust).
*** p<0.01.
** p<0.05.
[a] p<0.1.

daily fruit or vegetable consumption in the East-West dimension; and BMI, education and weekly sports activity in the South-West dimension) are substantially reduced after weighting with the inverse of the propensity score, which points to a reasonable post-estimation balance across the explanatory variables. In fact, for most variables, the absolute value of the inverse-propensity weighted standardized difference is below 0.10, the threshold of appropriate balance as suggested by Austin (2009).

Finally, Table C2 displays the results of the "best linear prediction" method of Chernozhukov et al. (2018a,b). For the East-West model, neither coefficients differ significantly from one, suggesting an appropriate fit both in terms of the average treatment effect and treatment effect heterogeneity. Meanwhile, for the South-West model, the coefficient of the average effect is essentially one, but the coefficient of effect heterogeneity takes an imprecisely estimated negative value (which does not differ significantly either from zero or one). In line with Fig. 3, this also suggests a more homogeneous treatment effect in the South-West dimension.

4.4. Management

Finally, we analyse how the new diagnosis of diabetes is associated with changes in dietary habits. Table 4 indicates that among the overweight or obese population in wave 4 who remained non-diabetic throughout waves 4 – 7, weight loss due to diet in wave 7 was less prevalent by 5.1 %points in Eastern Europe than in Western Europe (OR = 0.28), while there was no difference in the South-West dimension. Compared to this population, the diagnosis of diabetes increased the prevalence of weight loss in the West (by 7.7 %points) as well as in the East (by 3.8 %points), while there was no change in the South. These results are only slightly affected when control variables are included in the analysis, although then the South-West difference in the prevalence of weight loss in the non-diabetic (baseline) sample becomes statistically significantly negative (2.5 %points lower in the South than in the West). Hence, in Western as well as in Eastern Europe, a new diabetes diagnosis is associated with a substantially increased likelihood of weight loss due to diet but no such association is found in the South.

Looking at two-year transitions to diabetes and using data from waves 4 to 7 for the available countries, we observe that in the wave when diabetes is first reported, 11.9% of the respondents who were overweight or obese in the previous wave report weight loss due to diet. This ratio is around 7-8% two and four years before and two and four years after the diagnosis, which is close to the average probability of weight loss due to diet among the overweight or obese population (6.7%). Hence, the change in dietary habits mostly coincides with the diagnosis of diabetes and thus the analysis above merits attention.

5. Conclusions

Using data from three regions and 15 countries in Europe, we documented that diabetes prevalence and incidence are much higher in the South and East than in the West, and only 25 – 70 % of these differences disappear by controlling for individual-level demographic and socio-economic characteristics, health status and health behaviour. The country-specific indicators of health spending, availability of physicians and prevalence of unmet needs explain a large portion of the remaining part. Thus, the observed regional differences are likely to be caused by a combination of the differences in healthcare systems and in individual socio-economic and health-related variables.

Heterogeneity analyses showed that the East-West difference in incidence is essentially zero for the least vulnerable groups such as those with tertiary education or without hypertension. At the same time, Western European countries fare much better in preventing diabetes among lower-educated individuals and among those with comorbidities. Meanwhile, the South-West difference seems more stable across these dimensions.

Our results on the association of diabetes prevalence and incidence with regional, socio-economic, health and lifestyle indicators are in line with the existing literature (Agardh et al., 2011; Espelt et al., 2013; IDF, 2019; Diabetes Prevention Program Research Group, 2002; Narayan et al., 2007; Whiting et al., 2011). On the other hand, we extend our understanding of the regional differences in diabetes by showing that these differences are larger among the high-risk individuals and – at least for Eastern Europe – are essentially zero among the lowest-risk population.

Using an indicator of change in dietary habits, we also found that overweight or obese individuals are less likely to change diet effectively in the South and especially in the East than in the West. However, among people newly diagnosed with diabetes, the prevalence of weight loss due to diet is similar in the East and in the West. Thus, at least for the East, we do not see evidence that the higher incidence of diabetes would be coupled with worse management as measured by weight loss due to diet.

The analysis is subject to some limitations. First, it uses self-reported data on diagnosed diabetes, although undiagnosed cases make up one-third to one-half of total (diagnosed and undiagnosed) prevalence (IDF, 2019). The explanatory variables such as BMI, hypertension or high blood cholesterol are also self-reported and thus are subject to measurement error. Second, the three examined regions are not homogeneous, hence, by construction, any regional analysis overlooks the differences across countries within a region. (Meanwhile, the country-level sample sizes are generally too small to yield powerful conclusions.) Third, diabetes incidence is only analysed over a six-year horizon, due to the changing country-composition of the SHARE sample. Finally, only a crude and self-reported measure of diabetes management – weight loss due to diet – is used because we do not observe detailed outpatient, inpatient or laboratory testing data (apart from the raw number of doctoral visits).

In the 6th wave of SHARE, blood samples were collected from around 24 thousand individuals from 12 countries. Since the blood parameters (including glycated hemoglobin [HbA1c] measurements) are not yet available for researchers, it remains for future research to analyse the prevalence of undiagnosed diabetes and the quality of diabetes management based on blood sample results.

Overall, we found that diabetes incidence in Eastern and Southern Europe is more than twofold higher than in Western Europe, and these differences cannot be explained by differences in the demographic composition, education level and economic activity of the population. Our results suggest that to reduce the regional differences in diabetes incidence in Europe, more emphasis should be put on the prevention of diabetes among individuals more prone to the disease in Eastern and Southern Europe, which could at least partly be achieved by interventions aimed at preventing obesity, hypertension or high cholesterol among the high-risk population.

Funding

The authors were supported by the "Lendület" programme of the Hungarian Academy of Sciences (grant number: LP2018-2/2018). Péter Elek was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and by the ÚNKP-19-4 New National Excellence Programme of the Hungarian Ministry for Innovation and Technology.

Conflict of interest

None.

Acknowledgments

The authors would like to thank two anonymous referees, Noémi Kreif and participants at the annual conference of the Hungarian Society of Economists as well as at the annual conference of the Austrian Health Economics Association for useful comments.

Appendix A. Details of the causal forest methodology

A basic algorithm for estimating $\tau(x)$ based on equation (1) is the following (Procedure 1 in Wager and Athey (2018), called double-sample tree). A random subsample is chosen without replacement, and is split into two parts: one half will be used for partitioning the tree, and the other half for estimating the treatment effect within each leaf of a tree. Partitioning is done by maximising the variance of $\hat{\tau}(X_i)$ on the first sample, and the effects are estimated afterwards on the second sample. Random subsampling and tree building are then repeated many times and the resulting effect estimates are averaged.

It turns out that the above procedure works well for estimating heterogeneous treatment effects in a randomised setting but does poorly in the presence of confounding (Athey et al., 2019). Hence the causal forest algorithm as implemented within the R package *grf* makes some important changes as follows.

Motivated by the partialling-out interpretation of multivariate regression, the basic idea is that if $\tau = \tau(x)$ is constant then

$$\hat{\tau} = \frac{\sum_i \left(Y_i - \hat{m}^{(-i)}(X_i)\right)\left(W_i - \hat{p}^{(-i)}(X_i)\right)}{\sum_i \left(W_i - \hat{p}^{(-i)}(X_i)\right)^2} \tag{2}$$

is a semiparametrically efficient estimator of τ (see Athey and Wager, 2019), where $\hat{m}^{(-i)}$ and $\hat{p}^{(-i)}$ denote "out-of-bag" random forest estimates of the regression function $m(x)$ and the propensity score $p(x)$, respectively. ("Out-of-bag" means that the i-th observation is not used in the estimation.) Furthermore, a non-constant $\tau(x)$ can be estimated as

$$\hat{\tau}(x) = \frac{\sum_i \alpha_i(x)\left(Y_i - \hat{m}^{(-i)}(X_i)\right)\left(W_i - \hat{p}^{(-i)}(X_i)\right)}{\sum_i \alpha_i(x)\left(W_i - \hat{p}^{(-i)}(X_i)\right)^2}, \tag{3}$$

where $\alpha_i(x)$ is a data-based kernel that can be determined with a forest-based procedure.

More specifically, the algorithm proceeds as follows. First, the effect of X on Y and W are partialled out by conventional random forest predictions and subsequent steps are carried out on the orthogonalised data $\widetilde{Y}_i = Y_i - \hat{m}^{(-i)}(X_i)$ and $\widetilde{W}_i = W_i - \hat{p}^{(-i)}(X_i)$. Second, in the training phase, a forest is grown recursively, by maximising in each tree split the heterogeneity of the estimated treatment effects across the resulting child nodes. (The idea is similar to equation (1) but a numerical approximation is used to speed up computations.) Third, in the forecast phase, the $\alpha_i(x)$ values of equation (3) are calculated by gathering a weighted list of the sample neighbours that fall into the same tree leaf as x. This third step is similar to the weighting-based interpretation of conventional random forests. Indeed, beyond the usual "averaging-across-trees" interpretation, forecasts from random forests can also be viewed as $\hat{m}(x) = \sum_i \alpha_i(x) Y_i$, where $\alpha_i(x)$ is a data-adaptive kernel that measures how often X_i falls into the same final tree leaf as x and how large the corresponding tree leaf is (Athey et al., 2019).

The resulting causal forest can be used to estimate average treatment effects on various subsamples. A naive estimator would be the average of the $\hat{\tau}_i = \hat{\tau}(X_i)$ values taken over a particular subsample S, but, following Athey and Wager (2019), Farrell (2015) and using the built-in function of the *grf* package, this can be made more precise with an augmented inverse propensity weighting (AIPW) correction:

$$\hat{ATE} = \frac{1}{n}\sum_{i:X_i \in S}\left\{\hat{\tau}_i + \frac{W_i - \hat{p}_i}{\hat{p}_i}(1-\hat{p}_i)\left((Y_i - \hat{m}_i) - (W_i - \hat{p}_i)\hat{\tau}_i\right)\right\}, \tag{4}$$

where $\hat{p}_i$ and $\hat{m}_i$ are the estimates of the propensity score and the regression function, respectively, and n is the size of the subsample. This modification ensures that the estimator is doubly robust, i.e. it provides valid inference if either the propensity score function or the regression function (but not both of them) is misspecified.

Finally, the fit of the estimated causal forests should be evaluated. First, the overlap assumption can be checked by looking at whether the propensity scores are bounded away from zero and one. Second, covariate balance can be examined by comparing the inverse-propensity weighted averages of the explanatory variables across the two groups. (Here, treatment observations are weighted by $1/\hat{p}_i$ and control observations by $1/(1-\hat{p}_i)$.) Third, the "best linear predictor" method of Chernozhukov et al. (2018a,b) can be implemented. In this method, motivated by equation (2), $\tilde{Y}_i$ is regressed on $C_i = \overline{\tau}\widetilde{W}_i$ and $D_i = \left(\hat{\tau}^{(-i)}(X_i) - \overline{\tau}\right)\widetilde{W}_i$, where $\hat{\tau}^{(-i)}(X_i)$ is the out-of-bag treatment effect estimate and $\overline{\tau}$ is its sample average. If the coefficent of C_i is one then the model captures the average treatment effect adequately, and if the coefficient of D_i is one (or at least significantly positive) then the heterogeneity of the treatment effect is well calibrated, too. (See Athey and Wager, 2019 for more details.)

Appendix B. Comparison of causal forest and subsample-specific OLS results

Fig. B1

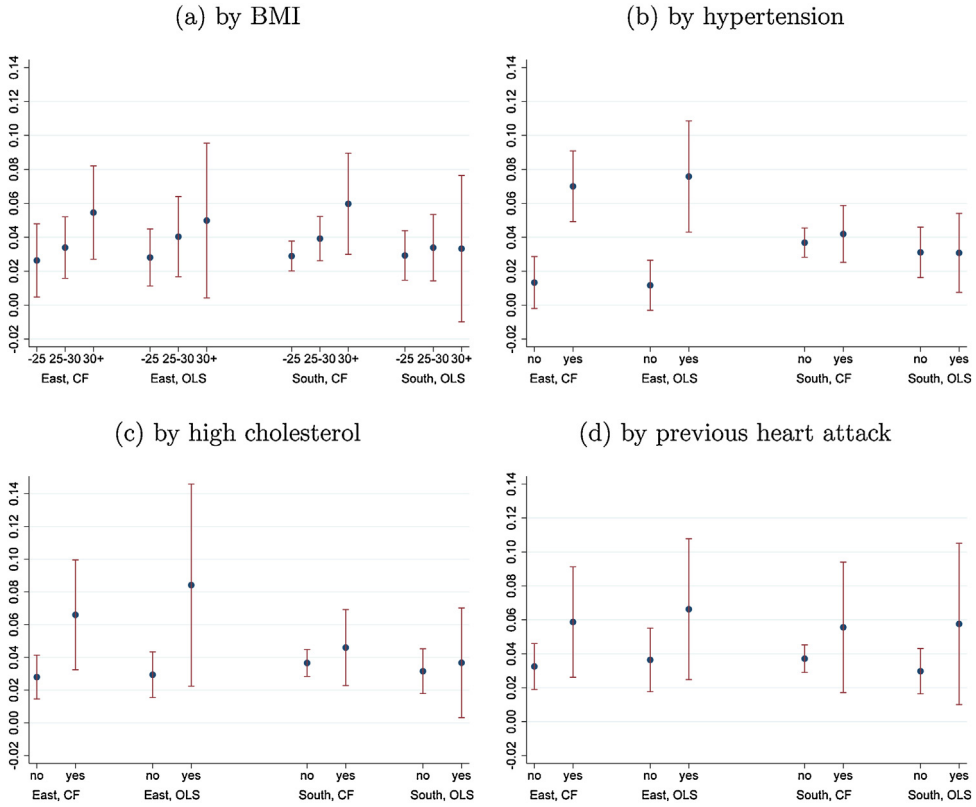

Fig. B1. Adjusted East-West and South-West differences based on the causal forest (CF) analysis and subsample-specific ordinary least squares (OLS) regressions, with 95% confidence intervals.

Appendix C. Goodness-of-fit analysis of the estimated causal forests

Fig. C1
Tables C1 and C2

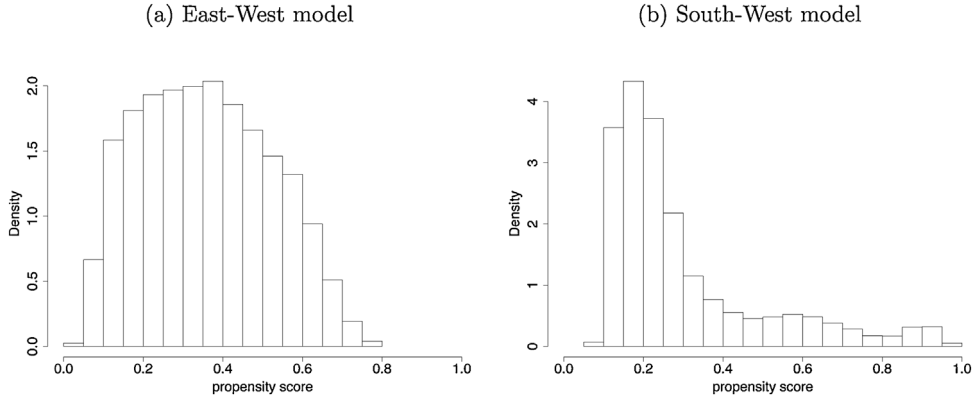

Fig. C1. Histogram of the propensity scores in the causal forest models.

Table C1
Raw and inverse-propensity weighted means of the explanatory variables in the East-West and the South-West causal forest models of diabetes incidence.

	Raw			Inverse-propensity weighted		
	means		stand. diff.	means		stand. diff
East-West comparison	East	West		East	West	
female	0.614	0.577	0.075	0.618	0.577	0.084
age	20.51	20.43	0.009	20.62	20.44	0.019
years of education	11.445	11.069	0.092	11.789	11.103	0.166
employment	0.290	0.353	−0.135	0.328	0.334	−0.012
BMI	27.58	25.98	0.356	26.39	26.55	−0.036
hypertension	0.436	0.333	0.213	0.373	0.363	0.021
high cholesterol	0.155	0.199	−0.117	0.145	0.203	−0.151
heart attack	0.157	0.084	0.226	0.122	0.095	0.088
stroke	0.040	0.025	0.084	0.035	0.027	0.046
smoker	0.206	0.181	0.063	0.184	0.191	−0.016
sports weekly	0.520	0.555	−0.071	0.548	0.538	0.020
fruit or vegetable daily	0.678	0.805	−0.293	0.785	0.748	0.088
South-West comparison	South	West		South	West	
female	0.573	0.577	−0.008	0.559	0.575	−0.031
age	21.04	20.43	0.066	20.58	20.62	−0.004
years of education	8.02	11.07	−0.652	10.06	10.24	−0.038
employment	0.245	0.353	−0.237	0.293	0.334	−0.088
BMI	26.84	25.98	0.202	26.42	26.15	0.063
hypertension	0.383	0.333	0.104	0.355	0.341	0.030
high cholesterol	0.203	0.199	0.010	0.200	0.202	−0.006
heart attack	0.078	0.084	−0.023	0.068	0.084	−0.063
stroke	0.021	0.025	−0.027	0.016	0.028	−0.080
smoker	0.162	0.181	−0.048	0.172	0.178	−0.017
sports weekly	0.400	0.555	−0.316	0.466	0.525	−0.119
fruit or vegetable daily	0.841	0.805	0.093	0.846	0.805	0.108

Standardized difference: difference of the means divided by the square root of the average of the two individual variances.

Table C2
Results from the "best linear predictor" method to evaluate the fit of the causal forests.

	East-West model			South-West model		
	coef.	S.E.	95% C.I.	coef.	S.E.	95% C.I.
mean forest prediction (coef. of C_i)	0.89	(0.17)	[0.55; 1.24]	0.95	(0.23)	[0.51; 1.39]
differential forest prediction (coef. of D_i)	1.49	(0.76)	[0.01; 2.97]	−0.87	(1.71)	[-4.22; 2.48]

References

Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T., Sidorchuk, A., 2011. Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. Int. J. Epidemiol. 40 (3), 804–818. doi:http://dx.doi.org/10.1093/ije/dyr029.

Athey, S., Wager, S., 2019. Estimating treatment effects with causal forests: an application. Observational Studies 5, 36–51 URL https://obsstudies.org/277-2/.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. Ann. Statist 47 (2), 1148–1178. doi:http://dx.doi.org/10.1214/18-AOS1709.

Austin, P.C., 2009. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. Commun. Stat. - Simul. Comput 38 (6), 1228–1234. doi:http://dx.doi.org/10.1080/03610910902859574.

Bashkin, O., Horne, R., Bridevaux, I.P., 2018. Influence of health status on the association between diabetes and depression among adults in Europe: findings from the SHARE international survey. Diabetes Spect. 31 (1), 75–82. doi:http://dx.doi.org/10.2337/ds16-0063.

Beilfuss, S.N., Thornton, J.A., 2016. Pathways and hidden benefits of healthcare spending growth in the US. Atl. Econ. J. 44 (3), 363–375. doi:http://dx.doi.org/10.1007/s11293-016-9506-6.

B“orsch-Supan, A., 2019. Survey of Health, Ageing and Retirement in Europe (SHARE) wave 7. Release version: 7.0.0.. doi:http://dx.doi.org/10.6103/SHARE.w7.700 Accessed: 2019-05-02.

B”orsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S., 2013. Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). Int. J. Epidemiol. 42 (4), 992–1001. doi:http://dx.doi.org/10.1093/ije/dyt088.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018a. Double/debiased machine learning for treatment and structural parameters. J. Econom 21 (1), C1–C68. doi:http://dx.doi.org/10.1111/ectj.12097.

Chernozhukov, V., Demirer, M., Duflo, E., Fernandez-Val, I., 2018b. Generic Machine Learning Inference On Heterogenous Treatment Effects In Randomized Experiments. Working Paper 24678. National Bureau of Economic Research URL https://www.nber.org/papers/w24678.

Diabetes Prevention Program Research Group, 2002. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl. J. Med 346 (6), 393–403. doi:http://dx.doi.org/10.1056/NEJMoa012512.

Espelt, A., Borrell, C., Palència, L., Goday, A., Spadea, T., Gnavi, R., Font-Ribera, L., Kunst, A.E., 2013. Socioeconomic inequalities in the incidence and prevalence of type 2 diabetes mellitus in Europe. Gac. Sanit. 27 (6), 494–501. doi:http://dx.doi.org/10.1016/j.gaceta.2013.03.002.

Eurostat, 2020. Eurostat Database, Population And Social Conditions, Health, Health Care. URL https://ec.europa.eu/eurostat/data/database (Accessed: 2020-04-04).

Farrell, M.H., 2015. Robust inference on average treatment effects with possibly more covariates than observations. J. Econom 189 (1), 1–23. doi:http://dx.doi.org/10.1016/j.jeconom.2015.06.017.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer Series in Statistics. Springer New York Inc, New York, NY, USA.

Heijink, R., Koolman, X., Westert, G.P., 2013. Spending more money, saving more lives? The relationship between avoidable mortality and healthcare spending in 14 countries. Eur. J. Health Econ. 14 (3), 527–538. doi:http://dx.doi.org/10.1007/s10198-012-0398-3.

Huang, I., Lim, M.A., Pranata, R., 2020. Diabetes mellitus is associated with increased mortality and severity of disease in COVID-19 pneumonia - a systematic review, meta-analysis, and meta-regression. Diabetes Metab. Syndr. 14 (4), 395–403. doi:http://dx.doi.org/10.1016/j.dsx.2020.04.018 ISSN 1871-4021..

IDF, 2019. Diabetes Atlas, 9th edition. https://diabetesatlas.org/en/resources/; Accessed: 2019-12-12.

Kassi, E., Pervanidou, P., Kaltsas, G., Chrousos, G., 2011. Metabolic syndrome: definitions and controversies. BMC Med. 9, 48. doi:http://dx.doi.org/10.1186/1741-7015-9-48.

Kouwenhoven-Pasmooij, T., Burdorf, A., Roos-Hesselink, J., Hunink, M., Robroek, S., 2016. Cardiovascular disease, diabetes and early exit from paid employment in Europe; the impact of work-related factors. Int. J. Cardiol. 215, 332–337. doi:http://dx.doi.org/10.1016/j.ijcard.2016.04.090 ISSN 0167-5273..

Li, S., Wang, J., Zhang, B., Li, X., Liu, Y., 2019. Diabetes mellitus and cause-specific mortality: a population-based study. Diabetes Metab. 43 (3), 319–341. doi:http://dx.doi.org/10.4093/dmj.2018.0060.
Malter, F., Börsch-Supan, A., 2015. SHARE wave 5: Innovations & Methodology. Technical Report. MEA at the Max Planck Institute for Social Law and Social Policy, Munich URL http://www.share-project.org/fileadmin/pdf_documentation/Method_vol5_31March2015.pdf.
Narayan, K.V., Boyle, J.P., Thompson, T.J., Gregg, E.W., Williamson, D.F., 2007. Effect of BMI on lifetime risk for diabetes in the US. Diabetes Care 30 (6), 1562–1566. doi:http://dx.doi.org/10.2337/dc06-2544.
Newhouse, J.P., 1992. Medical care costs: how much welfare loss? J. Econ. Perspect. 6 (3), 3–21. doi:http://dx.doi.org/10.1257/jep.6.3.3.
Rodriguez-Sanchez, B., Cantarero-Prieto, D., 2019. Socioeconomic differences in the associations between diabetes and hospital admission and mortality among older adults in Europe. Econ. Hum. Biol. 33, 89–100. doi:http://dx.doi.org/10.1016/j.ehb.2018.12.007.
Rumball-Smith, J., Barthold, D., Nandi, A., Heymann, J., 2014. Diabetes associated with early labor-force exit: a comparison of sixteen high-income countries. Health Aff. (Millwood) 33 (1), 110–115. doi:http://dx.doi.org/10.1377/hlthaff.2013.0518.
Tamayo, T., Rosenbauer, J., Wild, S., Spijkerman, A., Baan, C., Forouhi, N., Herder, C., Rathmann, W., 2014. Diabetes in Europe: an update. Diabetes Res. Clin. Pract. 103 (2), 206–217. doi:http://dx.doi.org/10.1016/j.diabres.2013.11.007.
Tibshirani, J., Athey, S., Wager, S., 2019. grf: Generalized Random Forests. URL https://CRAN.R-project.org/package=grf (R package version 0.10.4).
Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Am Stat. Assoc. 113 (523), 1228–1242. doi:http://dx.doi.org/10.1080/01621459.2017.1319839.
Whiting, D.R., Guariguata, L., Weil, C., Shaw, J., 2011. IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. Diabetes Res. Clin. Pract. 94 (3), 311–321. doi:http://dx.doi.org/10.1016/j.diabres.2011.10.029.
WHO, 2020. European Health Information Gateway, Physicians per 100000. URL https://gateway.euro.who.int/en/indicators/hfa_494-5250-physicians-per-100-000/ (Accessed: 2020-04-04.).