The impact of diabetes on labour market outcomes in Mexico: a panel data and biomarker analysis

January 3, 2019

Abstract

Recent evidence for Mexico suggests important differences in health status between people with diagnosed and undiagnosed diabetes. However, there is at best scarce evidence on the economic consequences of diabetes, especially in contexts where the condition often remains undiagnosed, as is typically the case in low- and middle income countries. Using Mexican longitudinal as well as biomarker data we estimated the impact of diabetes and diabetes duration on employment probabilities, wages and working hours. We further explored how these effects differed for those ~~aware and those unaware of their~~ with diagnosed and undiagnosed diabetes. For the longitudinal analyses nationally representative data from 11836 men and 13745 women 15 to 64 years old were taken from three waves (2002, 2005, 2009) of the Mexican Family Life Survey. We estimated a ~~within-between effects~~ fixed-effects model to account for unmeasured time-invariant confounders of diabetes~~while simultaneously assessing differences between subjects with and without diabetes. The within effects indicated~~ . We found a reduction in the probability of being employed between 5.4 and 6.0 percentage points for men and women, respectively, but no effects on hours worked or wages. ~~The between effects showed similar results.~~ Employment probabilities fell gradually with each year since diagnosis. Using cross-sectional biomarker data, we observed that 68% of those exhibiting glycated hemoglobin (HbA1c) levels above the clinical diabetes threshold did not self-report a diagnosis, hence were undiagnosed. Nevertheless, regression analysis revealed that there was no association of diabetes with labour outcomes for undiagnosed women or men. This suggests that results based on self-reported diabetes cannot be extended to the considerable population with undiagnosed diabetes, likely because of a selection of people in worse health and with a longer diabetes duration into the diagnosed population. An earlier diagnosis and improved treatment of diabetes therefore may prevent adverse health effects and related economic hardship in Mexico.

1. Introduction

Diabetes, a disease characterized by elevated blood glucose levels due to the body's inability to use insulin properly, has in the last two decades increasingly become a problem for low- and middle-income countries (LMICs) as well as high-income countries (HICs), with over two-thirds of people with diabetes living in the developing world (International Diabetes Federation 2015). In Mexico, diabetes prevalence is estimated to have grown from 6.7% in 1994 to 14.4% in 2006 (Barquera, Campos-Nonato, et al. 2013) and 15.8% in 2015. Diabetes has become the number one contributor to mortality (International Diabetes Federation 2015), by increasing the risk for heart disease and stroke, blindness, kidney disease and ~~nerve problems, food~~ neurologic problems, foot ulcers and amputations (Reynoso-Noverón et al. 2011). However, via effective self-management of the disease ~~,~~ ~~many if not all of the complications can be avoided~~ through regular monitoring, behaviour change and medication adherence, the occurrence of complications could be avoided or delayed in many cases (Gregg et al. 2012; Lim et al. 2011).

The observed increase in diabetes incidence has been attributed to a deterioration in diet and a reduction in physical activity (Barquera, Hernandez-Barrera, et al. 2008; Basu et al. 2013), while genetic predisposition among Mexicans with pre-Hispanic ancestry may also play a role (Williams et al. 2013). The onset of diabetes has been occurring at an ever earlier age in Mexico (Bello-Chavolla et al. 2017), increasing the risk of complications occurring during the productive lifespan. Only a minority of patients in Mexico achieves adequate blood glucose control (Barquera, Campos-Nonato, et al. 2013). Moreover, diabetes ~~in Mexico coexist with high levels of infectious diseases , exposing~~ is related to diseases such as depression, hypertension and cardiovascular disease that burden the health system ~~to a 'double-disease burden' increasing the pressure to identify treatment priorities and use existing resources efficiently~~(World Health Organization 2016).

Despite the catastrophic impact of diabetes on health, its economic consequences, in particular in LMICs have received less attention, especially its effects on labour outcomes

(Anonymous 2007). The latter have been studied predominantly in high-income countries, where substantial economic losses have been observed (Brown, Pagán, et al. 2005; Brown, Perez, et al. 2011; Brown 2014; Latif 2009; Minor 2011, 2013; Minor and MacEwan 2016). For LMICs less evidence is available. One study exploited a natural experiment in China and found a significant reduction in income due to a recent diabetes diagnosis (Liu and Zhu 2014). A study for Mexico, using cross-sectional data from 2005, found a significant ($p<0.01$) reduction in employment probabilities for males by 10 percentage points (p.p.) and for females by 4.5 p.p. ($p<0.1$) (Anonymous 2007). Most existing studies rely on instrumental variable (IV) estimation to address the potential endogeneity of diabetes using the genetic component of diabetes based on its family history as an instrument. However, family history of diabetes may also proxy for other genetically transferred traits, including unobserved abilities, as well as intrahousehold or intergenerational dynamics that impact labour outcomes directly; the validity of this IV therefore remains debatable. Panel data methods provide the opportunity to account for time-invariant unobserved individual characteristics, which may play an important role, but have not yet been used by our knowledge. Such unobservables, ~~like for instance health endowments~~ for instance hunger or nutrient deficiency experienced in early life, could adversely affect health as well as the propensity to develop type 2 diabetes ~~in particular~~ later in life (Ewijk 2011; Li et al. 2010; Sotomayor 2013)~~; they may also affect~~ . Additionally, there may also be long-term effects on labour outcomes—either directly through ~~their effects on~~ reductions in contemporaneous productivity (Currie and Vogl 2013), or indirectly by limiting educational attainment and human capital accumulation (Ayyagari et al. 2011). These unobservables thereby present a major source of a potential bias that can be accounted for by the use of panel data estimation.

In parallel to these identification challenges, heterogeneity in impact and measurement across the population also deserves further investigation. Recent evidence from Mexico points to a strong positive relationship of diabetes duration with mortality due to diabetes

related complications (Herrington et al. 2018). A longer disease duration was found to be related with higher glycated hemoglobin (HbA1c) levels and undiagnosed diabetes had the lowest diabetes related mortality risks. The latter points to potential selection issues when using self-reported diabetes data to investigate economic outcomes~~as those~~. Those who self-report, and hence tend to be diagnosed, are in worse health than those undiagnosed~~, potentially leading~~. This can lead to an overestimation of the economic effects of diabetes, in particular in populations with a large undiagnosed population, such as in many LMICs (Beagley et al. 2014). So far, however, little evidence exists on the economic impact according to diabetes severity~~and~~, duration or for those with undiagnosed diabetes.

The objective of this study was to provide new evidence on the impact of diabetes on labour outcomes, adding to previous work by paying close attention to the challenges of unobserved heterogeneity, to the chronic nature of diabetes and to ~~the population unaware of their condition (i.e. the 'undiagnosed')~~undiagnosed diabetes. We used three waves of ~~Mexican panel data~~the Mexican Family Life Survey (MxFLS), covering the period 2002–2012. Applying a ~~within-between model, which models individual fixed effects and between effects separately,~~ fixed effects model we accounted for time-invariant heterogeneity when assessing the impact of self-reported diabetes and time since diagnosis on labour outcomes. ~~We also used rich and novel~~ To assess the role uf undiagnosed diabetes we used biomarker data from the ~~most recent wave of data, we assessed the role of undiagnosed diabetes~~last wave of the MxFLS.

2. Data

This paper uses data from the Mexican Family Life Survey (MxFLS), a nationally representative longitudinal household survey, containing three waves conducted in 2002, 2005–2006 and 2009–2012. ~~Data was collected~~ It is the only longitudinal household survey in Mexico that provides data on a wide range of social, demographic, economic and health

characteristics (Rubalcava and Teruel 2013). Because the survey followed participants moving within Mexico as well as to the US, around 90% of the original sample have been reinterviewed in the third wave. Our samples were restricted to the working age population (15–64) and excluded pregnant women~~and those in school~~. Pregnant women have an increased diabetes risk and may not be able to work. Since their inclusion may have biased the estimates we dropped all observations of women reporting to be pregnant at the time of the survey (N=764). We also dropped those reporting to be in school. The first part of the analysis used all three waves and the panel structure of the data. The second part used a biomarker subsample of the third wave (2009–2012). Because the biomarker sample included everybody above the age of 44 but only a random subsample of those aged 44 or below (Crimmins et al. 2015), its age structure was older and hence its self-reported diabetes prevalence higher. The analysis therefore compares with self-reported data for this specific subsample only.

Our outcome variables of interest were employment status, weekly working hours, hourly wage, and occupation. Employment status was defined as having carried out an activity that helped with the household expenses the last week and working for at least four hours per week. We explicitly included informal employment and employment without monetary remuneration, for instance in family businesses. Hourly wage was constructed as reported monthly income from the first and second job, divided by average number of weeks per month and weekly working hours. Labour income was obtained from the response to questions on wages, income from piecework, tips, income from extra hours, meals, housing, transport, medical benefits and other earnings, or from the response to a question on aggregate labour income for the entire month. We adjusted calculated wages for inflation in the year of interview and considered the log of real wages. Due to a considerable number of missing or zero income reports, the sample used for the wage estimation was smaller than the sample for working hours. Working hours were combined from both the first and a potential second job. Descriptive statistics for the

entire panel sample show that 86% of men reported some form of employment compared to 37% of women (see Table 1). Interestingly, men did not report considerably higher hourly wages than women but worked more hours per week. Men also more often worked in agricultural jobs while women were more likely to be self-employed or in non-agricultural wage employment. The educational attainment of women was lower than that for men on average.

Table 1. **Descriptive statistics for the panel sample (2002,2005,2009).**

	Males			Females		
	No diabetes	Diabetes	p (t-test)	No diabetes	Diabetes	p (t-test)
Dependent variables						
Employed	0.87	0.80	0.00	0.37	0.26	0.00
Hourly wage (in Mexican Peso)	42.29	46.79	0.83	40.67	36.33	0.61
Weekly working hours	46.83	46.51	0.60	39.06	37.51	0.09
Non-agricultural worker or employee	0.51	0.41	0.00	0.24	0.13	0.00
Agricultural worker	0.19	0.13	0.00	0.02	0.01	0.00
Self-employed	0.16	0.26	0.00	0.09	0.11	0.04
Diabetes variables						
Diabetes duration (years)		7.40			7.79	
Control variables						
Age	35.31	50.68	0.00	35.37	50.45	0.00
Any medical insurance	0.47	0.59	0.00	0.50	0.62	0.00
City of 2,5oo-15,000	0.11	0.09	0.03	0.11	0.13	0.00
City of 15,000-100,000	0.10	0.14	0.00	0.10	0.10	0.40
City of >100,000	0.34	0.39	0.00	0.35	0.34	0.47
Married	0.53	0.77	0.00	0.53	0.66	0.00
Number of children (age<6) in household	1.49	1.14	0.00	1.60	1.13	0.00
Indigenous group	0.19	0.15	0.00	0.19	0.19	0.86
Education						
Secondary	0.31	0.22	0.00	0.31	0.16	0.00
High school	0.16	0.07	0.00	0.14	0.03	0.00
Higher education	0.11	0.12	0.39	0.10	0.03	0.00
Wealth index	0.00	0.04	0.27	−0.01	0.01	0.36
N	20391	994		25664	1666	

Notes Mean values. Diabetes refers to self-reported diabetes.

The first part of the analysis focused on the relationship of labour outcomes with self-reported diabetes, which was based on the survey question: "Have you ever been diagnosed with diabetes?". Because the data did not distinguish between type 1 and type 2 diabetes, we assumed that the estimates represented the impact of type 2 diabetes, by far the most common type of diabetes in Mexico. As a robustness check, we re-estimated our main results categorizing diabetes into early-onset and late-onset cases, according to

the age at which diabetes was first reported in the survey. This was a similar approach to Alegre-Díaz et al. (2016), who assumed that everybody diagnosed before age 35 and using insulin had type 1 diabetes. Accordingly, we assumed that those first reporting a diabetes diagnosis before the cut-off had type 1 diabetes while those above had type 2 diabetes. Nonetheless, because we cannot warranty that this is 100% accurate as it may be unlikely that both populations consisted exclusively of one type of diabetes, we preferred to think of the groups as of early- and ~~late onset groups(in the case of the within-coefficient, which only takes into account incident cases). Because we did not have~~ late-onset groups. This separation also provides information about the ~~exact age at diagnosis for all diabetes cases in all three waves,~~ effects for different age groups, as the late-onset group had an average age of onset of 50 compared to 28 for the ~~between coefficient may also stratify people with diabetes into the late onset group , even though they actually had early onset diabetes but only joined the sample after already having been diagnosed for several years~~early-onset group. In the pooled data, which combines all three waves, diabetes was self-reported by 5% of men and 6% of women~~, respectively~~. This is consistent with other reports from Mexico for the time, showing a prevalence of diagnosed diabetes of 7.5% in 2006 in a sample also including people over the age of 64 (Barquera, Campos-Nonato, et al. 2013). Apart from self-reported diabetes, which was available in all rounds, we also used information on the self-reported year of diagnosis as well as biometrically measured HbA1c levels for a subsample of respondents from the third wave.

Information on the self-reported year of diagnosis reported in the third wave allowed us to construct a measure of time since diagnosis. For those also present in previous waves, we inferred the time since diagnosis by the difference between the year of the interview and the year of diagnosis. This allowed us to use panel data methods for the duration analysis as well, however limited to those reporting the year of diagnosis in the third wave.

The second part of the analysis assessed the role of ~~measurement error associated with self-reported diabetes. This was done by also considering those with undiagnosed diabetes,~~

~~i.e. the false negatives~~undiagnosed diabetes. The biometrically measured blood glucose value that allowed us to identify those with undiagnosed diabetes, was available for over 6000 respondents in the third wave. We chose the internationally recognized cut-off of an HbA1c $\geq$ 6.5% to define diabetes as recommended by the World Health Organization (WHO) (World Health Organization 2011). As we show in Supplementary Table S6, several observations of self-reported diabetes had HbA1c levels below the diabetes threshold. We dropped those for our analysis as it was unclear if they had misreported their diabetes status or had achieved these low levels as a result of their successful disease management. Analysis including those cases did not lead to qualitatively different results (results available on request).

3. Estimation strategy

To investigate the relationship between self-reported diabetes and three labour outcomes—employment, wages and weekly working hours—we estimated a ~~effects model, an extension of the correlated random effects modelfirst proposed by Mundlak (1978), that explicitly models both within and between effects . The within effects are identical to a~~ fixed effects model. The fixed effects (FE) model ~~, accounting~~ accounts for the potential bias introduced by time-invariant unobservables, providing an estimate of the effect for cases that received a diagnosis throughout the survey~~(705 incident cases compared to 970 non-changing diabetes cases in the used sample). Modeling the between effect allowed us to also use information from those that already had diabetes at baseline~~.

$$Y_{it} = \beta_0 + \beta_1(D_{it} - \overline{D}_i) + \beta_2(X_{it} - \overline{X}_i) + \beta_3\overline{D}_i + \beta_4\overline{X}_i + (u_i + e_{it}), \quad (1)$$

The ~~within-effect is arrived at by~~ fixed effects model uses only the within-person variation for identification, i.e. the difference between the diabetes indicator D_{it} and its cluster mean $\overline{D}_i$, so that β_1 represented the within-person variation of diabetes over time. The

same applies to the other time-varying covariates X_{it}. ~~Cluster means of diabetes and of all other time-varying covariates were also included to capture the between effect. The error terms u_i and e_{it} capture the errors for the within and between variation, respectively.~~ Y_{it} was a binary variable taking a value of 1 if respondent i reported being in employment at time t and 0 otherwise. ~~Making use of the user written Stata command xthybrid , we estimated effects applying a multilevel mixed-effects generalized linear model.~~ For ease of interpretation we chose to estimate a linear probability model for the effects of diabetes on employment.

~~For the outcomes~~ To estimate the effect on working hours and wages, our empirical models were estimated conditional on being in employment. Y_{it} represented the log hourly wage or the weekly working hours over the last year, for respondent i at time t.

~~The regressions controlled for~~ Because the fixed effects model accounts for any time-invariant confounding, we only included time-variant confounders. We controlled for changes in the level of urbanization, ~~education, state~~the educational level, the state of residence, marital status, the number of children below the age of 6 in the household, a quadratic age term ~~,~~ and calendar year dummies~~as well as household wealth~~ . We also control for household wealth approximated by a household asset index. We composed an indicator using principal component of household assets and housing following Filmer et al. (2001) (Filmer and Pritchett 2001). The ~~assets indicators~~ asset index reflected owning a vehicle, a second house, a washing machine, dryer, stove, refrigerator or furniture, any electric appliances, any domestic appliances, a bicycle, farm animals, and accounted for the physical condition of the house, proxied by the type of floor material and water access. In our main regression models we did not account for body mass index (BMI). While part of the effect of diabetes may be due to potential adverse effects of obesity, including BMI as a control variable in the model would have led to biased estimates if the diagnosis of diabetes also had an effect on BMI, which ~~was likely~~ has been shown to be the case ~~. In general, control variables should not also be potential outcome variables , hence we~~

~~similarly~~ in other studies (De Fine Olivarius et al. 2015; Seuring, Suhrcke, et al. 2018; Slade 2012). Similarly, we did not control for ~~other chronic diseases that may have been caused by~~ any diseases that are likely consequences of diabetes, such as hypertension or cardiovascular disease, as this would prevent any causal interpretation of the relationship of diabetes with our labour market outcomes (Angrist and Pischke 2009). Stata 15 was used for all analyses (StataCorp 2017).

3.1. Labour outcomes and time since diagnosis

The chronic nature and irreversibility of diabetes provide good reason to explore the long term effects post diagnosis. To do this we ~~estimated the following model:~~

$$Y_{it} = \beta_0 + \beta_1(Dyears_{it} - \overline{Dyears}_i) + \beta_2(X_{it} - \overline{X}_i) + \beta_3\overline{Dyears}_i + \beta_4\overline{X}_i + (u_i + e_{it}),$$

~~where $\beta_1 Dyears_{it}$ was continuous~~ replaced the binary diabetes indicator of Eq 1 with a continuous variable indicating years since the diagnosis was first reported. While simultaneous inclusion of year dummies and time since diagnosis (which varies by one unit in each time period) would typically not allow separate identification of the coefficient of time since diagnosis ~~in Eq 3.1. and Eq **??**,~~ identification here relied on the presence of people without diabetes in the sample, for which diabetes duration did not increase.

We also considered a spline function that allowed for non-linear effects over time~~.~~

$$Y_{it} = \beta_0 + \beta_1(Dsplines_{it} - \overline{Dsplines}_i) + \beta_2(X_{it} - \overline{X}_i) + \beta_3\overline{Dsplines}_i + \beta_4\overline{X}_i + (u_i + e_{it}),$$

, with $g(Dyears_{it}) = \sum_{n=1}^{N} \delta_n \cdot max\{Dyears_{it} - \eta_{n-1}\}I_{in}$ and $I_{in} = 1[\eta_{n-1} \leq Dyears_{it} < \eta_n]$, with η_n being the place of the n-th node for $n = 1, 2, \ldots, N$. The coefficient δ_n captures the effect of diabetes for the n-th interval. The effects are linear if $\delta_1 = \delta_2 =, \ldots, = \delta_n$.

Based on visual inspection (Fig. 1 on page 17) we chose three nodes located at 3, 7 and 12 years after diagnosis. The first three years should capture any immediate effects of the diagnosis, the years four to seven any effects during time of adaptation to the disease and the later terms the long term effects. We also estimated a non-linear model using dummy variables for duration groups rather than splines, applying the same duration cut-offs. Because the year of diagnosis was only reported in the third wave, time since diagnosis was not available for those who were not interviewed in the third round. A reported diagnosis in the year of the interview was counted as 'one year since diagnosis'.

3.2. Labour outcomes and biometrically measured diabetes

The biomarker analysis consisted of three steps. We first re-estimated Eq 2 to assess the relationship between self-reported diabetes with labour outcomes, but this time for the cross-sectional biomarker sample only, using the following specification:

$$Y_i = \beta_0 + \beta_1 Dsr_i + \beta_2 X_i + c_i + v_i \tag{2}$$

where v_i were community fixed effects which reflected local unobserved characteristics, such as access to healthcare, poverty and unemployment in the community. Communities (or *localidades* in Spanish) are the smallest administrative units (nested within municipalities) recognized by the Mexican Institute for Statistics and Geography (INEGI). We did not use household fixed effects since the average number of observations per household was close to one.

In a second step we estimated the relations between biomarker diabetes and labour outcomes, using the following equation:

$$Y_i = \beta_0 + \beta_1 Dbio^{d}_{\_i} + \beta_2 X_i + v_i + u_i, \tag{3}$$

where ~~$Dbio^{d}$~~ $Dbio_i$ was equal to 1 if HbA1c ~~$\geq 6.5\%$~~ $\geq 6.5\%$.

To estimate the effect of undiagnosed diabetes, we added self-reported diabetes ~~and interacted it with the biomarker~~ back into the equation (Eq 4).

$$Y_i = \beta_0 + \beta_1 Dsr_i + \beta_2 Dbio_i + \beta_3 Dsr_i * Dbio_i + \beta_4 X^d{}_i + v_i + u_i. \quad (4)$$

~~Note that the interaction term~~ This changes the interpretation of ~~β_1 and β_2, with β_1 now representing the effect on those aware of their condition but with levels below the diabetes threshold; while~~ β_2 which now reflects the effect on those with undiagnosed diabetes, i.e. the respondents not self-reporting diabetes but with HbA1c levels equal to or above the threshold. ~~The interaction term β_3 shows the effect for those with self-reported diabetes and levels above the threshold.~~

We further investigated the effect of the severity of diabetes on labour outcomes, replacing $Dbio^d$ with $Dbio^c$, a variable that was 0 for HbA1c $< 6.5\%$ and took the actual value of HbA1c minus 6.4 for those with an $HbA1c \geq 6.5\%$ (Eq 5). This allowed us to investigate the effect of a one percentage point increase in HbA1c levels for people with undiagnosed diabetes (β_2) as well as for those with self-reported diabetes ~~above the diabetes threshold~~ (β_3).

$$Y_i = \beta_0 + \beta_1 Dsr_i + \beta_2 Dbio^c_i + \beta_3 Dsr_i * HbA1c_i + \beta_4 X_i + v_i + u_i. \quad (5)$$

4. Results

4.1. Labour outcomes and self-reported diabetes

The results of estimating Eq 1 in Table 2 indicated significant and substantial reductions in the probability of employment for men and women with self-reported diabetes. ~~The overall similarity of between estimates suggests that the within effect is generalizable to the entire self-reporting diabetes population, i.e. not only for representative of those that~~

~~developed diabetes after joining the survey. Additionally, it provides suggestive evidence that time-invariant unmeasured confounders may play a limited role.~~ Employment probabilities were reduced by over ~~5~~ 5.4 p.p. for ~~both genders~~men and 6 p.p. for women, translating into relative reductions of ~~14% for women and~~ of 6% for men and 14% for women. There was no significant relationship between diabetes ~~on one hand and working hours and wageson the other, though the between estimates suggested that men with diabetes earned generally more than their counterparts without diabetes (Table 2)~~with working hours or wages. Overall these results ~~thus~~ suggested effects at the extensive margin (employment), but not at the intensive margin (labour supply and productivity).

Dividing the diabetes population into early and late onset groups, men, and potentially also women, saw their employment probabilities negatively affected by a diabetes onset later in life (Supplementary Table S2). ~~For women, a particularly strong effect was also found for early diabetes onset. The between estimator suggested that men and women with diabetes were less likely to be employed at older ages, but not at a younger age~~In particular, women with an early diabetes onset experienced an adverse effect. For working hours, ~~we only found an adverse effect using the within effects estimator for early onset, which may have been spurious due to low incidence rates of diabetes. For wages~~effects are less precise but may indicate increased working hours for men with an early diabetes onset, while women reduced their work hours. Finally, we found ~~a positive effect of diabetes incidence on women. However, the within estimates for early onset cases again may be spurious. The between estimates show that especially older menwith diabetes received higher wages than those without diabetes~~higher wages for women with an early diabetes onset, but no effects for men.

To assess whether diabetes affected the selection into different types of work, we investigated the role of diabetes for the probability of being in non-agricultural wage employment, agricultural employment or self-employment. ~~The within effects estimator showed~~ We found a reduction in the probability to work in agriculture for women, ~~while we~~

Table 2. **Labour outcomes and self-reported diabetes**

	Employment		Weekly work hours		Log hourly wages	
	Males	Females	Males	Females	Males	Females
Diabetes ~~(within)~~	−0.054**	~~-0.060~~-0.059**	~~-0.582~~−0.506	~~-1.990~~−1.998	~~0.063~~0.055	~~0.074~~0.081
	(0.025)	(0.024)	(~~1.501) (0.848~~1.499)	(~~1.306~~2.511)	(~~0.046~~0.068)	(~~0.064~~0.158)
~~Within–Between (p value)~~N	21388	27339	17616	9112	13828	7068

Notes Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children < 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

~~found no statistically significant effects for men . The between effects suggested that men with diabetes were less likely to be employed in agriculture, but more likely to be self-employed. Women with diabetes were less likely to be employed in agricultural and non-agricultural jobs~~ but not for men (Table 3). Disaggregating the diabetes groups further according to their age showed that most statistically significant relationships were driven by the ~~older onset~~ older-onset group (Supplementary Table S3). ~~Interestingly, for~~ For male self-employment, ~~incidence of~~ diabetes increased the probabilities to be self-employed in the younger group, while it reduced the probabilities to be self-employed in the ~~older onset group. However, especially the results for early onset diabetes in women should be interpreted carefully due to limited number of diabetes cases.~~ older-onset group.

Table 3. **Selection into types of work and self-reported diabetes.**

	Males			Females		
	Non-agric.	Agric.	Self-employed	Non-agric.	Agric.	Self-employed
Diabetes ~~(within)~~	~~-0.007~~−0.006	−0.008	~~-0.042~~−0.043	−0.001	−0.022**	~~-0.030~~-0.029
	(0.029)	(0.022)	(0.026)	(0.018)	(0.009)	(0.018)
N	20719	20719	20719	26575	26575	26575

Notes Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children < 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Because obesity is one of the main risk factors for developing diabetes and may also affect labour market outcomes, its exclusion from above models may lead to biased estimates. We therefore reestimated all regressions in this section including a binary control

for obesity (BMI $\geq$ 30) (Supplementary Table S7 and Table S8). This led to a reduction in sample size due to the larger number of missing cases for BMI. Obesity itself did not appear to be an independent predictor of any labour market outcome. The estimates of the effect of diabetes remained similar for most outcomes. Only for male employment probabilities, the diabetes coefficient was no longer statistically significant at the 10 percent significance level. Estimating the original model without accounting for obesity but using the sample with only non-missing BMI cases showed similar changes in effects, suggesting that these changes were rather the result of a smaller sample size than of the inclusion of obesity.

4.2. Labour outcomes and time since diagnosis

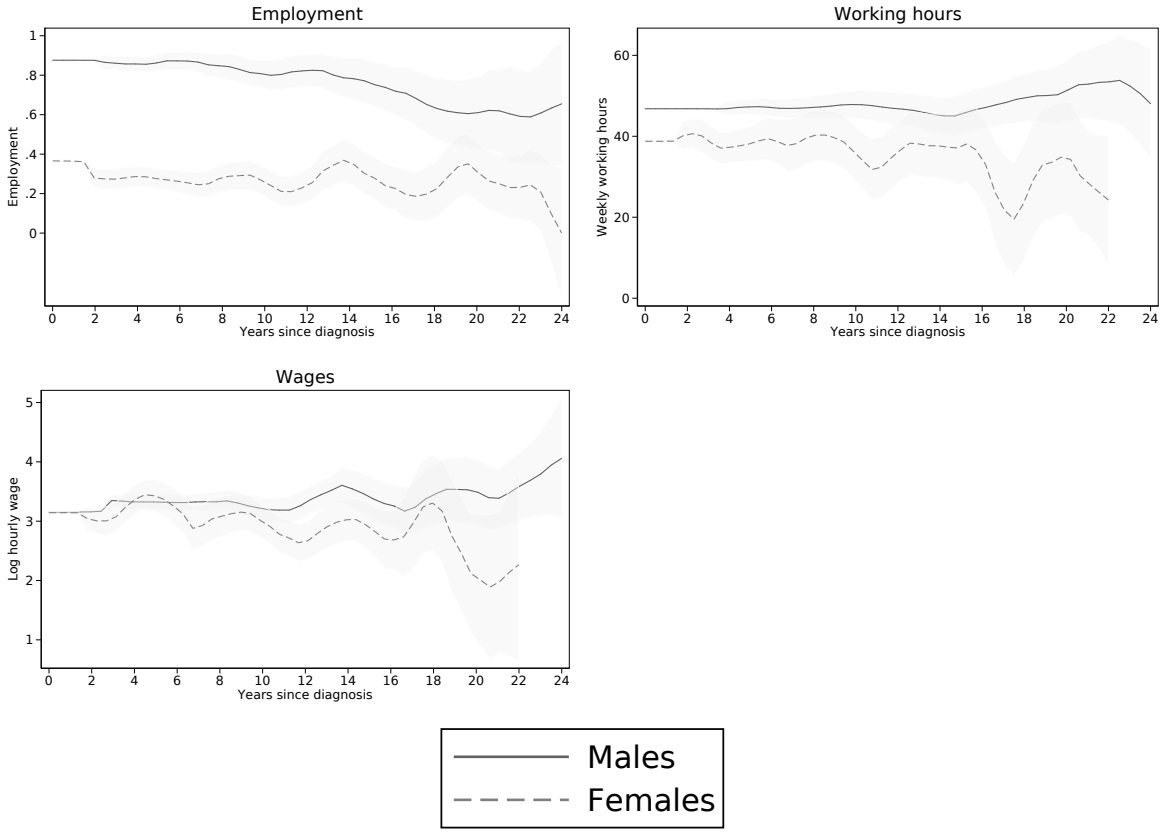
Fig. 1 shows that the probability of employment for men steadily declined as time progressed, using a non-parametric kernel-weighted local polynomial regression. For women, a first drop-off occurred right after diagnosis; though no consistent pattern emerged thereafter. The dynamics for working hours and wages were less clear, with a possibly long term negative trend for women but not for men.

Table 4 panel A shows the results of estimating Eq 3.1., which indicated that male and female employment probabilities fell every year, with a larger effect shown by the within coefficient.

~~For females, the within coefficient also suggested a reduction in wages~~

Wages were reduced for women as diabetes progressed ~~while the between coefficient showed no association for women and a relatively small positive association with male wages~~using the linear specification. Using diabetes onset groups, there was no evidence of an effect of diabetes duration for early onset groups (see Supplementary Table S4). ~~However, again the within results for early onset should be interpreted with caution due to the~~ Unfortunately the very limited number of diabetes ~~incidence cases in this group , which also~~ cases in the early-onset group prohibited the estimation of ~~any within~~ effects of

Figure 1: Employment, wages, working hours and years since self reported diabetes: Kernel-weighted local polynomial regression



Notes The shaded areas indicate the 95% confidence intervals.

early diabetes onset duration ~~for~~ on wages and working hours. ~~The results also indicated that the effects found in Table 4 were driven mainly by those with a diabetes onset after age 35.~~

The non-linear results for the spline function and dummy variable approach are presented in panels B and C, respectively. They suggest that the main adverse effects appeared after a prolonged time of living with diabetes; i.e. after more than seven years since diagnosis. The same was true for female wages. The lack of a statistically significant effect for the earlier years of diabetes duration may have been due to a reduction in statistical efficiency ~~, reduced by~~ due to the sole reliance on within-variation and the

Table 4. **Relationship between self-reported years since diagnosis and employment probabilities using continuous duration and duration splines.**

	Employment		Weekly work hours		Log hourly wages	
	Males	Females	Males	Females	Males	Females
Panel A: linear effect						
Years since diagnosis ~~(within)~~	−0.016***	−0.009*	0.185	0.115	−0.016	−0.067**
	(0.006)	(0.005)	(0.334)	(0.652)	(0.018)	(0.029)
Panel B: splines						
Years since SR diagnosis						
0–3	−0.013	−0.018	0.708	2.953	−0.005	0.047
	(0.014)	(0.016)	(0.857)	(2.700)	(0.054)	(0.124)
4–7	−0.011	−0.002	0.215	−2.517	−0.032	−0.131
	(0.014)	(0.014)	(0.761)	(1.752)	(0.046)	(0.101)
8–12	0.003	−0.003	−1.153	1.144	−0.009	−0.053
	(0.021)	(0.014)	(1.252)	(1.635)	(0.065)	(0.061)
13+	−0.039***	−0.015	0.720	0.184	−0.007	−0.096***
	(0.014)	(0.010)	(0.943)	(1.414)	(0.057)	(0.037)
Panel C: dummies						
0–3	0.005	−0.007	0.352	17.309*	0.223	−0.447
	(0.052)	(0.059)	(3.123)	(9.975)	(0.186)	(0.549)
4–7	−0.031	−0.049	2.860	10.878	0.047	−0.568
	(0.042)	(0.050)	(2.664)	(9.504)	(0.127)	(0.544)
8–12	−0.066	−0.026	−0.709	13.733	−0.133	−0.873*
	(0.063)	(0.059)	(4.181)	(9.695)	(0.207)	(0.521)
13+	−0.134	−0.062	−3.379	13.309	0.164	−0.882**
	(0.098)	(0.068)	(4.715)	(9.239)	(0.284)	(0.446)
N	16298	22427	10771	5746	13583	7391

Notes Panel A presents the results of the linear specifications. Panel B presents the results of the non-linear specifications. Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children < 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

separation into duration groups~~and into within and between variation~~. Re-estimating the specifications with a random effects model ~~that combined both types of variation into one estimate~~ showed that, at least for the models with dummies, ~~there was a more or less~~ an immediate reduction in employment probabilities that became stronger the longer a person had diabetes (see Supplementary Table S5). Note that we did not estimate models splitting diabetes in early and late onset groups, as this implied strong reductions in statistical power.

To test the robustness of the results to controlling for obesity, we estimated above models including a dummy variable for obesity (Supplementary Table S9). Overall, results remained very similar and obesity itself was not found to be affecting any labour market outcome.

4.3. Cross-sectional biomarker analysis

As reported in Supplementary Table S6, 18% of the observations in the biomarker sample were ~~false negatives, i.e.~~ undiagnosed. Further, 2% ~~were false positives, though the latter may have included cases that received a diabetes diagnosis and managed to reduce their to non-diabetes levels via medication and~~/reported diabetes but had an HbA1c below the threshold. Because we did not know if the latter were true diabetes cases that had been able to return to non-diabetic levels as a result of their treatment regime or lifestyle changes (Flores-Hernández et al. 2015)~~. Overall~~, or if they represented misreports, we dropped these observations from the sample. Overall, 80% of the self-reports ~~were consistent with the biomarker data~~also had diabetes according to their HbA1c levels. Comparing the health status and diabetes risk factors of the diagnosed and undiagnosed diabetes populations suggested that those with self-reported diabetes were older and in worse health, both objectively and subjectively compared to those undiagnosed. This suggests a selection into the diagnosed group based on the severity and potentially duration of diabetes. If the adverse effects of diabetes are due to its health impact, we would suspect worse labour market outcomes for the diagnosed compared to the undiagnosed population.

Table 6 presents the results from estimating Eq. 2 – 5. Panel A confirms the earlier longitudinal results using self-reported diabetes for the cross-sectional biomarker sample. The results in panel B indicate that the relationship with employment became weaker when using diabetes defined by the biomarker instead of self-reported diabetes, in particular for men. Results in Panel C were obtained from estimating Eq. 4 and indicate an absence of a (statistically significant) negative relationship between undiagnosed diabetes ~~(expressed in the 'Biomarker diabetes but not self-reported' coefficient)~~ and labour outcomes. ~~The coefficients for the interaction term were mostly negative, though only statistically significant in the case of female working hours.~~

To explore whether the adverse effects increased with higher HbA1c levels, we estimated Eq. 5. The results in panel D ~~again only suggest a negative association of a one~~ only

Table 5. **Descriptive comparison of diagnosed and undiagnosed population with diabetes.**

	Males			Females		
	Diagnosed diabetes	Undiagnosed diabetes	P value (t-test)	Diagnosed diabetes	Undiagnosed diabetes	P value (t-test)
Employed	0.807	0.875	0.012	0.241	0.331	0.002
Hourly wage	35.931	30.670	0.120	36.092	32.638	0.550
Usual weekly working hours	44.341	46.682	0.104	34.708	39.681	0.046
Age	53.162	44.720	0.000	53.167	44.449	0.000
Any medical insurance	0.677	0.599	0.033	0.733	0.643	0.002
City of 2,5oo-15,000	0.096	0.107	0.643	0.112	0.112	0.994
City of 15,000-100,000	0.135	0.096	0.098	0.087	0.090	0.884
City of >100,000	0.331	0.297	0.325	0.294	0.333	0.185
Married	0.719	0.643	0.031	0.633	0.569	0.035
Number of children (age<6) in household	0.950	1.122	0.073	0.960	1.257	0.001
Indigenous group	0.158	0.211	0.079	0.195	0.207	0.626
Primary	0.479	0.434	0.238	0.626	0.465	0.000
Secondary	0.212	0.231	0.554	0.132	0.233	0.000
High school	0.062	0.131	0.003	0.037	0.117	0.000
Higher education	0.139	0.113	0.288	0.030	0.073	0.003
Wealth index	−0.174	0.117	0.000	0.012	0.103	0.157
Subjective health						
very good	0.023	0.094	0.000	0.012	0.054	0.001
good	0.212	0.434	0.000	0.180	0.367	0.000
fair	0.619	0.442	0.000	0.643	0.528	0.000
bad	0.135	0.026	0.000	0.155	0.048	0.000
very bad	0.012	0.004	0.187	0.010	0.004	0.246
Glycated hemoglobin (HbA1c)	9.037	8.533	0.004	8.979	8.680	0.049
Hypertension (self-reported)	0.262	0.074	0.000	0.397	0.150	0.000
Blood pressure						
Systolic	136.688	130.506	0.000	136.070	122.835	0.000
Diastolic	84.677	82.063	0.003	84.495	79.689	0.000
Heart disease (self-reported)	0.035	0.007	0.004	0.050	0.024	0.021
BMI	28.868	28.311	0.135	30.640	29.778	0.032
Obese (BMI $\geq$ 30)	0.338	0.311	0.440	0.469	0.431	0.225

were

Notes Mean values. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

show a borderline statistically significant adverse association of 0.9 percentage points per percentage point increase in HbA1c ~~with female working hours for those with self-reported diabetes. For undiagnosed diabetes, we again found no effects~~for female employment probabilities.

Table 6. **Biomarker results.**

	Employment		Weekly working hours		Log hourly wages	
	Males	Females	Males	Females	Males	Females
Panel A: Diabetes (self-reported)						
Self-reported diabetes	~~-.055~~-.057**	~~-.050~~.057**	~~-.308~~-.543	~~-.323~~-2.154	~~-.026~~-.057	−.005
	(~~-.026~~.025)	(~~-.024~~.026)	(~~-1.321~~1.427)	(~~-2.190~~2.433)	(~~-.066~~.070)	(~~-.105~~.121)
Panel B: Diabetes (biomarker)						
Biomarker diabetes (HbA1c ≥ 6.5)	~~-.012~~-.013	~~-.032~~.034*	~~-.060~~.018	~~1.067~~1.382	~~-.008~~-.005	~~-.041~~-.045
	(.016)	(.018)	(~~-.344~~.849)	(~~-1.470~~1.480)	(.045)	(.071)
Panel C: Self-reported and undiagnosed diabetes						
Self-reported diabetes ~~but tested negative~~ (β_1)	~~-.049~~-.061**	~~-.027~~-.042	~~1.547~~-.715	~~7.417~~-3.954	~~.250~~-.067	~~-.025~~.034
	(~~-.059~~.028)	(~~-.051~~.031)	(~~-3.446~~1.574)	(~~-4.669~~2.823)	(~~-.205~~.085)	(~~-.232~~.137)
~~Biomarker diabetes but not self-reported~~ Undiagnosed diabetes (HbA1c ≥ 6.5) (β_2)	~~.005~~-.006	−.020	~~.174~~-.224	~~2.304~~-2.394	~~.015~~-.014	~~-.049~~-.053
	(.018)	(.020)	(~~-.960~~.962)	(~~-1.639~~1.647)	(.050)	(.078)
Panel D: HbA1c levels						
Self-reported diabetes	~~-.067~~-.080*	~~-.035~~-.066	~~1.471~~-.084	~~4.974~~-4.463	~~.157~~-.061	~~-.017~~.011
	(~~-.072~~.046)	(~~-.041~~.046)	(~~-2.619~~2.409)	(~~-3.654~~1.592)	(~~-.119~~.107)	(~~-.186~~.227)
HbA1c if ≥ 6.5	~~.001~~-.005	~~-.003~~-.009*	~~-.005~~-.150	~~.235~~-.318	~~.002~~-.004	−.005
	(~~-.002~~.005)	(~~-.002~~.006)	(~~-.104~~.253)	(~~-.193~~.463)	(~~-.005~~.014)	(~~-.008~~.019)
Self-reported diabetes × HbA1c if ≥ 6.5	~~.001~~-.003	~~.000~~-.010	~~-.207~~-.064	~~-.852~~.375	~~-.023~~-.002	~~.005~~-.000
	(~~-.007~~.012)	(~~-.005~~.012)	(~~-.294~~.668)	(~~-.424~~1.043)	(~~-.014~~.030)	(~~-.022~~.052)
N	2749	3537	2276	1121	1787	866

Notes Results are based on community level fixed effects. Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children < 6, wealth, health insurance status, age squared and one dummy variable for each calendar year to account for the multiple years of data collection for the third wave. The wage and working hour models additionally control for type of work (agricultural and self employed with non-agricultural wage employment as the base).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5. Discussion

Diabetes is now one of the most common chronic diseases in low- and middle income countries, as well as high-income countries, with potential severe impacts on the health and economic well-being of those affected. Yet rigorous evidence on the economic consequences for LMICs remains scarce.

To address key methodological challenges, this paper used rich longitudinal panel data from Mexico that also contained diabetes biomarkers. The biomarker data showed alarming levels of clinically tested diabetes (27% prevalence) and indicate that a large proportion of the Mexican population (18%) ~~is unaware of their condition~~has undiagnosed diabetes.

~~Overall, the paper found~~ The paper provides evidence for adverse effects of self-reported diabetes on ~~the probability of being employed, confirming earlier findings for Mexico that used~~ employment, working hours and wages. While earlier evidence for Mexico already exists for employment (Seuring, Goryakin, et al. 2015), this was the first time effects on working hours and wages were investigated. Furthermore, we added to the study of Seuring, Goryakin, et al. (2015)by using longitudinal instead of cross-sectional ~~information. But, these new results also suggest a comparatively larger~~ data to identify a causal relationship. We provided first evidence of the long term impact of diabetes ~~on female employmentprobabilities. The evidence also points towards the main effects being~~ and

explored the extend and effects of undiagnosed diabetes. We confirmed the findings of Seuring, Goryakin, et al. (2015)insofar that we found an adverse effect of diabetes on male employment. We further show more conclusive evidence that also women experience a reduction in their employment probabilities due to diabetes. We also found that the effects were mainly driven by those with a diabetes onset at a relatively later state, consisting ~~most likely of people with~~ of older people with most likely type 2 diabetes. This was also found by Seuring, Goryakin, et al. (2015), albeit they had not considered the onset of diabetes but only stratified their sample into an older and younger age group. Analyses of the long term impact indicated that the employment probability fell gradually in the years following diagnosis. The results ~~for the~~ using a non-linear ~~models~~ model were less clear, potentially due to reductions in statistical power, but suggest that adverse effects became stronger with time since diagnosis. The linear effect contrasts with estimates for the USA, where such an effect ~~seemed to be absent,~~ was absent; however allowing for non-linearity revealed falling employment probabilities after 11 to 15 years for females and after 2-5 years for males (Minor 2013). ~~For working hours and wages, our results were more ambiguous with adverse effects being observed for time since diagnosis on female wages only. In summary, most of the adverse effects of diabetes were found at the extensive margin, mainly affecting the probability of employment, rather than the intensive margin, i.e. working hours or wages.~~

Overall, a relationship of diabetes with working hours or wages is mostly absent~~, except for the between estimator in the case of wages for men. Contrasting the between and within results suggests that this significant result may have arisen from individual selection~~. Although any explanation at this point is speculative, it may be that higher paid ~~men, who also tend to be more educated~~ and more educated individuals, were able to remain employed without experiencing wage reductions, for instance due to their particular set of skills. They may also have had access to better health care leading to better diabetes related health outcomes. Low paid workers, on the other hand, may lack access to quality

diabetes care, making it more likely that they develop severe complications earlier (Flores-Hernández et al. 2015). They are also more likely to be in informal employment and low skilled jobs, with less job security and are thus more prone to being laid off to be replaced with healthier workers.

Because self-reported diabetes may not adequately represent the entire diabetes population if the share of undiagnosed diabetes is relatively large and those undiagnosed differ from those diagnosed, estimates based on self-reported diabetes are likely to be biased. Our results using the cross-sectional biomarker data suggest indeed that those with undiagnosed diabetes were significantly healthier and younger. Consequently, diabetes based on biomarkers was found to be less related to reduced employment, compared to self-reported diabetes, in particular for men. Further analysis showed that this was due to the absence of any association between undiagnosed diabetes and employment. These results are similar to those found for the USA, where no statistically significant relationship was observed between undiagnosed diabetes on employment, while a significant effect of diagnosed diabetes was observed (Minor and MacEwan 2016). Our results further indicate that the difference in the employment effects of diagnosed and undiagnosed diabetes was not mediated by current HbA1c levels, as even undiagnosed cases with relatively high HbA1c levels did not show any adverse associations with labour market outcomes. This is similar to findings for Mexican-Americans in the USA, where employment outcomes were unrelated to higher HbA1c levels (Brown, Perez, et al. 2011). This may stem from HbA1c levels primarily being informative for the last three months, and not being the only indicator for the severity of diabetes. Overall, it seems that general health differences related to a longer diabetes duration and selection into the diagnosed population, for instance based on emerging diabetes related health problems, could be driving the adverse economic effects among those with a self-reported diagnosis.

Our study had several limitations. While ~~the within-coefficient~~ our model accounts for any time-invariant confounding, the estimates may have been affected by unobserved time-

variant confounders. Reverse causality, where employment status affects the propensity to develop or be diagnosed with diabetes, may also play a role. Existing studies that looked at this particular direction of causality, however, have not found strong evidence for an effect of employment status on diabetes (Bergemann et al. 2011; Schaller and Stevens 2015), though they were carried out in high-income countries. We do not control for the effects of obesity, hypertension, self-reported health or other diseases in our models due to the high probability that they are affected by diabetes themselves, which would prevent a causal interpretation of the estimates. Nonetheless, robustness checks including obesity indicate that our main findings remain unchanged. For the duration analysis an additional limitation, imposed by the data, was that the year of diagnosis was only reported in the third wave. While this still allowed us to construct an estimate of the time since diagnosis for the previous waves, it restricted the analysis to those that were present in the last wave, thereby excluding those that dropped out of the sample prior to the third wave. Finally, we used a WHO recommended HbA1c cut-off to diagnose diabetes, due to the lack of a Mexico specific cut-off. There is some evidence that HbA1c may be affected by ethnicity (Sacks 2011). Hence, if Mexican ethnicity would lead to different HbA1c levels, the use of our cut-off could have led to misclassifications based on the used biomarkers.

Despite these limitations, our findings bear important implications. First, the impact of self-reported diabetes on labor outcomes in Mexico seems mostly limited to its effect on employment probabilities, though there is some indication that it could also reduce wages over time for women. Second, its effect on employment was much stronger for females, though the underlying reasons for this remain unclear. Potential explanations are that lower working hours or wages for women make a dropout less costly. Other evidence suggests that women with diabetes are in worse metabolic health compared to men when they cross the diabetes threshold (Peters et al. 2015), making it more likely for them to drop out. Third, caution is needed when estimates based on self-reported diabetes are interpreted in terms of the entire population, i.e. extending to those with

undiagnosed diabetes. Ideally, studies would include a biomarker analysis, acknowledge the differences between diagnosed and undiagnosed ~~subpopulations~~sub-populations, and carry out a separate analysis whenever feasible. If this is not possible, study conclusions about the effect of self-reported diabetes should be limited to this specific part of the population, in particular in environments where the share of undiagnosed diabetes is high, as is the case for most low- and middle income countries.

~~While we find no effect for undiagnosed diabetes in our cross section analysis, further inquiry over time is needed.~~ The large proportion of previously undiagnosed cases found in this paper indicates that diagnosis—at least in Mexico—happens late or not at all. This may reduce the possibilities to prevent complications via treatment and self-management, increasing the risk of severe complications appearing premature. Earlier diagnosis and ensuing effective treatment may lighten the health and economic burden. ~~Further analysis is also needed to explain why the adverse economic effects are so large for women. Ultimately, prevention of diabetes is of high importance. Taxation of sugar sweetened beverages may be one promising way forward , though their long-term effectiveness remains unclear. Given the established links between early life heath and later life incidence~~ Therefore more research is needed to investigate the economic impact of diabetes over time. Longitudinal biomarker information could be used to observe the true duration and severity of diabetes as well as ~~other chronic diseases , investments in maternal and child health may be particularly attractive to address non-communicable and communicable diseases at the same time~~the time that passes till a diagnosis through a doctor. This would allow for a better understanding of when adverse economic effects start to appear. Further, future research should investigate how the time of diagnosis and the treatment of diabetes affect the occurrence of adverse labour market effects of diabetes. The results of such research could inform costing studies to include more detailed information on the indirect costs of diabetes; or cost-effectiveness analyses that aim to include a measure of the potential benefit of the intervention to employers or society at large.

Supplementary material

Strategies to deal with inconsistent self-reporting over time

Reporting error can pose a considerable challenge in the use of self-reported data. Fortunately, the MxFLS data provide several possibilities to assess the amount of misreporting and apply corrections before estimating the labour market effects of diabetes. In what follows we describe how we have dealt with inconsistencies in self-reported diabetes over time.

Throughout the surveys, self-reported diabetes was measured by the question 'Have you ever been diagnosed by diabetes'. One of the key advantages of panel data is the repeated measurement which results in more than one data point, allowing to uncover inconsistencies for cases with multiple observations. Very little is known about inconsistencies in self-reported diabetes over time. Zajacova et al. (2010) assess the consistency of a self-reported cancer diagnosis over time in the USA. The study found that 30% of those who had reported a cancer diagnosis at an earlier point failed to report the diagnosis at a later point in time. A more recent diagnosis was found to be reported with greater consistency possibly due to increasing recall problems as time since diagnosis advanced.

When assessing the MxFLS, we also found inconsistencies in the diabetes self-reports across the three waves, with between 10–20% of those reporting diabetes in one wave not doing so in one of the subsequent waves. To improve the validity of diabetes self-reports, we were interested in reducing the amount of reporting inconsistencies.

For diabetes, the main concern with mismeasurement is related to ~~false negatives. False positives~~ a lack of a diagnosis. Wrong self-reports indicating a diagnosis of diabetes are deemed less of a problem since incentives to report diabetes when one does not have it seem to be very limited—although we cannot exclude this. A study from China finds that the vast majority (98%) of those who self-report diabetes are tested positive for diabetes, while only a minority of those who are tested positive for diabetes (40%) actually self-

report the disease (Yuan et al. 2015). Our data showed a similar pattern, with a low proportion (3%) of the respondents being tested negative while self-reporting diabetes, while the majority of those who are tested positive (68%) do not self-report diabetes.

We used the above information to infer the "true" diabetes status for those with inconsistent reports. For respondents present in all three waves, we corrected inconsistencies as reported in Supplementary Table S1. We assumed that if diabetes was reported only once in the first two waves (either in 2002 or 2005) and then not reported again in the ensuing waves, this diabetes report was likely to be false (see lines 3 and 4 in Supplementary Table S1) and that the person never had received a diagnosis. If a diabetes diagnosis was however reported in two of the three waves (in 2002 and 2009 but not 2005, or in 2002 and 2005 but not in 2009) we assumed that the respondent had diabetes in all three waves (see lines 1 and 2 in Supplementary Table S1). For cases where we only had information from two waves, we assumed that if a diabetes diagnosis had been reported in a prior wave they also had diabetes in the ensuing wave, even if it was not reported in the latter (see lines 5 and 6 in Supplementary Table S1), given that most diabetes self-reports tend to be correct.

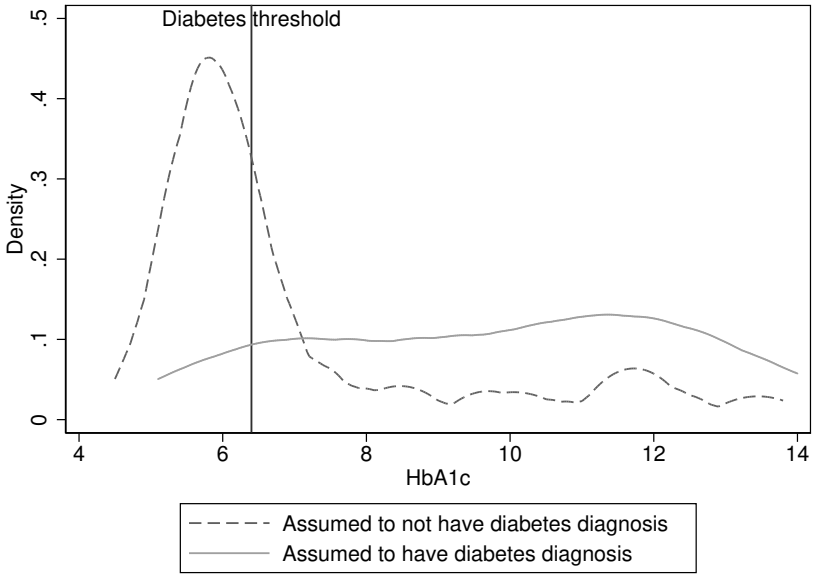
Table S1. Inconsistencies in diabetes self-report in MxFLS.

	Inconsistency	Assumption	Number of observations replaced
1	Diabetes self-report only in 2002, but not in 2005 and 2009	Has no diabetes in 2002 either	66
2	Diabetes self-report only in 2005, but not in 2002 and 2009	Has no diabetes in 2005 either	52
3	Diabetes self-report in 2002, 2005 but not in 2009	Has diabetes in 2009 as well	19
4	Diabetes self-report in 2002, 2009 but not in 2005	Has diabetes in 2005 as well	63
5	Diabetes self-report in 2002, but not in 2005. Not in survey in 2009	Has diabetes in 2005 as well	44
6	Diabetes self-report in 2005, but not in 2009. Not in survey in 2002	Has diabetes in 2009 as well	23

We then tested if those respondents we categorized as not having a diabetes diagnosis based on above rules were actually more likely to not have biometrically measured diabetes, using the biomarker data from wave 3. Of those with inconsistencies in their diabetes self-reports, 95 were present in the biomarker sample (46 with two self-reports (from lines 3 and 4 in Table S1) and 49 with one self-report of diabetes (from lines 1 and 2 in Supplementary Table S1)). Supplementary Figure S1 illustrates the difference between both groups and

suggests that indeed those with two self-reports of diabetes are much more likely to have HbA1c values above the diabetes threshold. A t-test comparing the mean HbA1c for the two groups indicates that those with two self-reports also have significantly ($p<0.001$) higher HbA1c levels than those with only one self-report of diabetes (9.7% vs. 7.1%). Further, of those with one self-report, only 30% have an HbA1c$\geq$ 6.5% compared to 87% of those with two self-reports. Based on these results it appears that we did minimize misclassification of people into diabetes or no-diabetes.

Figure S1: Kernel density of HbA1c values for those with one inconsistent and two inconsistent reports.



Early versus late onset of diabetes

Table S2. **Labour outcomes and self-reported diabetes by diabetes onset.**

	Employment		Weekly work hours		Log hourly wages	
	Males	Females	Males	Females	Males	Females
Early onset (within)	0.133	−0.206**	14.712*	−18.636*	−0.523	0.388***
	(0.176)	(0.086)	(8.370)	(9.668)	(0.340)	(0.057)
Late onset (within)	−0.059**	−0.048*	−1.008	−1.322	0.079	0.067
	(0.025)	(0.025)	(1.513)	(2.553)	(0.067)	(0.163)
Early onset (between)	0.016	−0.052	1.483	−2.640	−0.102	0.286
	(0.037)	(0.047)	(2.581)	(4.034)	(0.099)	(0.232)
Late onset (between)	−0.087***	−0.067***	−1.077	−0.869	0.121**	−0.111*
	(0.017)	(0.016)	(0.895)	(1.375)	(0.050)	(0.066)
N	21388	27339	13828	7068	17616	9112

Notes Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children < 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S3. **Selection into types of work and self-reported diabetes by diabetes onset.**

	Non-agric.		Agriculture		Self-employed	
	Males	Females	Males	Females	Males	Females
Early onset ~~(within)~~	0.030	−0.105	~~-0.225~~ -0.226	−0.068	0.328**	−0.027
	(0.216)	(0.074)	(0.139)	(0.047)	(0.161)	(0.048)
Late onset ~~(within)~~	~~-0.008~~ -0.007	0.007	~~-0.002~~ -0.001	~~-0.019~~-0.018**	~~-0.053~~-0.054**	~~-0.030~~ -0.029
	(0.029)	(~~0.018~~0.019)	(0.022)	(0.009)	(0.026)	(0.019)
N	20719	26575	20719	26575	20719	26575

Notes Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children < 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S4. **Relationship between self-reported years since diagnosis and employment probabilities using continuous duration by diabetes onset.**

	Employment		Monthly work hours		Log hourly wages	
	Males	Females	Males	Females	Males	Females
Early onset (within)	−0.009	0.002				
	(0.017)	(0.013)				
Late onset (within)	−0.012**	−0.007*	0.086	0.308	−0.006	−0.059***
	(0.005)	(0.004)	(0.275)	(0.504)	(0.013)	(0.022)
Early onset (between)	0.011***	−0.016***	−0.016	0.029	0.187	−0.401
	(0.004)	(0.005)	(0.012)	(0.051)	(0.333)	(0.896)
Late onset (between)	−0.009***	−0.005***	−0.060	−0.144	0.011**	−0.010
	(0.002)	(0.002)	(0.117)	(0.129)	(0.006)	(0.009)
N	16308	22450	13592	7394	10778	5748

Notes The within estimator for the effects of early onset diabetes on wages and working hours could not be estimates due to no within-variation for diabetes. Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children < 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Random effects model

Table S5. **Relationship between self-reported years since diagnosis and employment probabilities using continuous duration and duration splines (random effects).**

	Employment		Weekly work hours		Log hourly wages	
	Males	Females	Males	Females	Males	Females
Panel A: linear effect						
Years since diagnosis	−0.007***	−0.004***	0.039	−0.130	0.010**	−0.009
	(0.002)	(0.001)	(0.102)	(0.127)	(0.005)	(0.008)
Panel B: splines						
Years since SR diagnosis						
0–3	−0.008	−0.015**	−0.035	0.507	0.038**	0.034
	(0.006)	(0.006)	(0.346)	(0.614)	(0.017)	(0.029)
4–7	0.001	0.004	0.242	−0.570	−0.032	−0.048
	(0.011)	(0.011)	(0.665)	(1.062)	(0.032)	(0.052)
8–12	−0.008	0.002	−0.116	−0.080	−0.003	−0.074
	(0.015)	(0.011)	(0.855)	(1.098)	(0.041)	(0.050)
13+	−0.012	−0.004	0.035	−0.339	0.029	0.011
	(0.008)	(0.003)	(0.410)	(0.241)	(0.018)	(0.017)
Panel C: dummies						
0–3	−0.036*	−0.041**	−0.821	1.091	0.134**	0.021
	(0.021)	(0.021)	(1.154)	(1.826)	(0.054)	(0.083)
4–7	−0.014	−0.056**	0.877	1.200	0.093	−0.003
	(0.022)	(0.023)	(1.375)	(2.530)	(0.059)	(0.118)
8–12	−0.069*	−0.043	0.427	0.302	−0.070	−0.148
	(0.037)	(0.030)	(2.288)	(2.995)	(0.101)	(0.117)
13+	−0.121***	−0.043	−0.568	−2.104	0.242*	−0.279*
	(0.045)	(0.031)	(2.280)	(3.088)	(0.126)	(0.153)
N	16308	22450	13592	7394	10778	5748

Notes Panel A presents the results of the linear specifications. Panel B presents the results of the non-linear specifications. Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children < 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S6. **Number of observations with diabetes (HbA1c $\geq$ 6.5%) and self-reported diabetes.**

	$HbA1c < 6.5\%$	HbA1c $\geq$ 6.5%	Total
No self-reported diabetes (N)	4544	1181	5725
Row %	79%	21%	100%
Cell %	**71%**	**18%**	-
Self-reported diabetes (N)	129	554	683
Row %	19%	81%	100%
Cell %	**2%**	**9%**	-
Total (N)	4673	1735	6408

Robustness checks

Table S7. **Labour outcomes and self-reported diabetes controlling for obesity**

	Employment		Weekly work hours		Log hourly wages	
	Males	Females	Males	Females	Males	Females
Obese (BMI $\geq$ 30)	0.007	-0.005	-0.127	-1.144	0.018	0.082
	(0.012)	(0.013)	(0.773)	(1.188)	(0.038)	(0.061)
Diabetes	-0.046	-0.064**	-0.689	-0.169	0.036	0.033
	(0.028)	(0.027)	(1.772)	(2.904)	(0.078)	(0.183)
N	17992	24145	14866	7929	11711	6166

Notes Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children $<$ 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S8. **Selection into types of work and self-reported diabetes controlling for obesity.**

	Males			Females		
	Non-agric.	Agric.	Self-employed	Non-agric.	Agric.	Self-employed
Obese (BMI $\geq$ 30)	0.005	-0.032**	0.036***	-0.021*	0.003	0.010
	(0.017)	(0.013)	(0.014)	(0.011)	(0.004)	(0.009)
Diabetes	0.010	0.002	-0.060**	-0.011	-0.020**	-0.025
	(0.033)	(0.023)	(0.028)	(0.020)	(0.010)	(0.021)
N	17414	17414	17414	23458	23458	23458

Notes Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children $<$ 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S9. **Relationship between self-reported years since diagnosis and employment probabilities using continuous duration and duration splines and controlling for obesity.**

	Employment		Weekly work hours		Log hourly wages	
	Males	Females	Males	Females	Males	Females
Panel A: linear effect						
Obese (BMI $\geq$ 30)	0.003	-0.010	0.059	-0.412	0.026	0.035
	(0.012)	(0.014)	(0.831)	(1.247)	(0.040)	(0.064)
Years since diagnosis	-0.019***	-0.008	0.259	-0.008	-0.016	-0.073**
	(0.006)	(0.006)	(0.375)	(0.721)	(0.019)	(0.034)
Panel B: splines						
Years since SR diagnosis						
Obese (BMI $\geq$ 30)	0.003	-0.009	0.073	-0.371	0.027	0.036
	(0.013)	(0.014)	(0.832)	(1.247)	(0.040)	(0.064)
0–3	-0.014	-0.022	0.806	3.762	-0.070	0.015
	(0.015)	(0.017)	(1.051)	(3.169)	(0.057)	(0.139)
4–7	-0.003	0.009	-0.293	-3.921**	0.035	-0.121
	(0.018)	(0.015)	(0.914)	(1.811)	(0.044)	(0.108)
8–12	-0.023	0.001	-0.098	3.082*	-0.062	-0.085
	(0.022)	(0.016)	(1.350)	(1.736)	(0.066)	(0.074)
13+	-0.038**	-0.024*	0.855	-1.128	0.005	-0.065
	(0.017)	(0.012)	(1.029)	(1.421)	(0.053)	(0.063)
Panel C: dummies						
Obese (BMI $\geq$ 30)	0.005	-0.009	0.044	-0.378	0.026	0.031
	(0.012)	(0.014)	(0.831)	(1.245)	(0.040)	(0.064)
0–3	0.028	-0.032	1.484	22.434*	0.047	-0.658
	(0.059)	(0.065)	(3.825)	(11.579)	(0.212)	(0.622)
4–7	0.001	-0.054	2.399	12.909	0.013	-0.793
	(0.044)	(0.055)	(3.181)	(11.063)	(0.154)	(0.616)
8–12	-0.064	0.010	0.296	15.604	-0.293	-1.125*
	(0.069)	(0.066)	(4.994)	(11.038)	(0.247)	(0.583)
13+	-0.208**	-0.073	-1.966	17.459*	0.168	-1.090**
	(0.105)	(0.081)	(4.975)	(10.262)	(0.256)	(0.499)
N	13912	19972	11622	6487	9262	5054

Notes Panel A presents the results of the linear specifications. Panel B presents the results of the non-linear specifications. Robust standard errors in parentheses. All models include variables for states, urbanization, level of education, marital status, number of children < 6, wealth, health insurance status, age squared and one dummy variable for each calendar year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.