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 a lack of a diagnosis. Wrong self-reports indicating a diagnosis of diabetes we deemed less of a problem since incentives to falsely report a diabetes diagnosis seem to be very limited—although we cannot exclude this. A study from China found that the vast majority (98%) of those who self-reported diabetes were tested positive for diabetes, while only a minority of those who were tested positive for diabetes (40%) actually self-reported the disease (Yuan et al.

2015). Our data showed a similar pattern, with a low proportion (2%) of the respondents being tested negative while self-reporting diabetes, while the majority of those who were tested positive (68%) did not self-report diabetes.

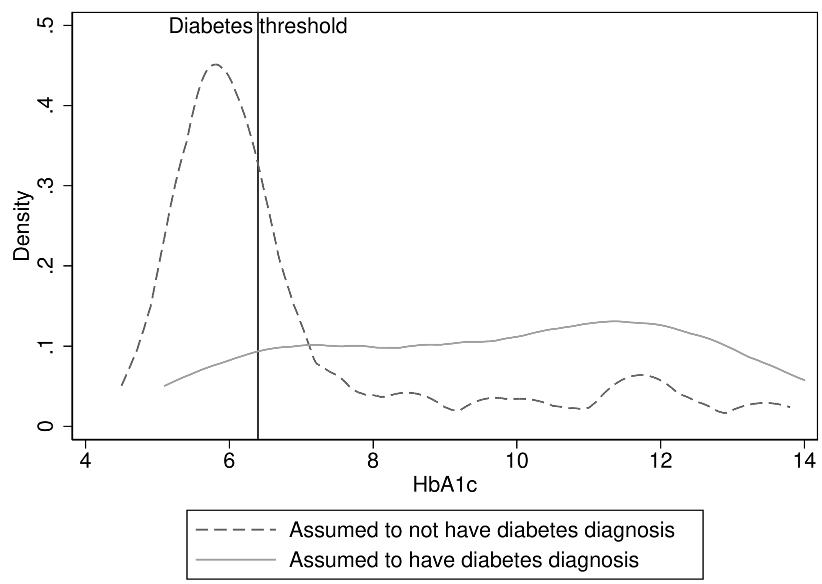
We used the above information to infer the "true" diabetes status for those with inconsistent reports. For respondents present in all three waves, we corrected inconsistencies as reported in Supplementary Table S1. We assumed that if diabetes was reported only once in the first two waves (either in 2002 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 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 the 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 were much more likely to have HbA1c values above the diabetes threshold. A t-test comparing the mean HbA1c for the two groups indicated that those with two self-reports also had 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% had 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 working hours		Log hourly wages	
	Males	Females	Males	Females	Males	Females
Early onset	0.133	-0.206^{**}	14.856^{*}	-18.250^{*}	-0.490	0.375^{***}
	(0.176)	(0.086)	(8.329)	(9.515)	(0.337)	(0.057)
Late onset	-0.059^{**}	-0.047^{*}	-0.936	-1.346	0.071	0.074
	(0.026)	(0.025)	(1.510)	(2.553)	(0.068)	(0.161)
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	0.030	−0.105	−0.226	−0.068	0.328**	−0.027
	(0.216)	(0.074)	(0.139)	(0.047)	(0.161)	(0.048)
Late onset	−0.007	0.007	−0.001	−0.018**	−0.054**	−0.029
	(0.029)	(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	-0.005	0.007				
	(0.017)	(0.013)				
Late onset	-0.012^{**}	-0.008^{**}	0.088	0.303	-0.007	-0.059^{***}
	(0.005)	(0.004)	(0.275)	(0.505)	(0.013)	(0.022)
N	16308	22450	13592	7394	10778	5748

Notes The effects of early onset diabetes on wages and working hours could not be estimated 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 working 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 working 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 working 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.