Volume Title: Insights in the Economics of Aging

Volume Author/Editor: David A. Wise, editor

Volume Publisher: University of Chicago Press

Volume ISBNs: 0-226-42667-X; 978-0-226-42667-9 (cloth);
978-0-226-42670-9 (e-ISBN)

Volume URL: http://www.nber.org/books/wise-21

Conference Date: April 30-May 2, 2015

Publication Date: March 2017

Chapter Title: Measuring Disease Prevalence in Surveys: A Comparison of Diabetes Self-Reports, Biomarkers, and Linked Insurance Claims

Chapter Author(s): Florian Heiss, Daniel McFadden, Joachim Winter, Amelie Wuppermann, Yaoyao Zhu

Chapter URL: http://www.nber.org/chapters/c13637

Chapter pages in book: (p. 227 – 252)

7

Measuring Disease Prevalence in Surveys
A Comparison of Diabetes Self-Reports, Biomarkers, and Linked Insurance Claims

Florian Heiss, Daniel McFadden, Joachim Winter, Amelie Wuppermann, and Yaoyao Zhu

7.1 Introduction

Reliable measures of disease prevalence are crucial for answering many empirical research questions in health economics, including the causal structures underlying the correlation between health and wealth. Much of the existing literature on the health-wealth nexus relies on survey data (for example, those from the US Health and Retirement Study [HRS]). Such survey data typically contain self-reported measures of disease prevalence, which are known to suffer from reporting error. Two more recent developments—the collection of biomarkers and the linkage with data from administrative sources such as insurance claims—promise more reliable measures of disease prevalence. In this chapter, we systematically compare these three measures of disease prevalence.

Florian Heiss is the Chair of Statistics and Econometrics at the University of Dusseldorf. Daniel McFadden is the Presidential Professor of Health Economics at the University of Southern California. He has joint appointments at the USC Sol Price School of Public Policy and the Department of Economics at USC Dornsife College. He is also a research associate of the National Bureau of Economic Research. Joachim Winter is the Chair of Empirical Economic Research at the University of Munich. Amelie Wuppermann is assistant professor of economics at the University of Munich. Yaoyao Zhu is a PhD candidate in economics at the University of Southern California.

This research was funded by the Behavioral and Social Research program of the National Institute on Aging (grants P01AG05842-18, P01AG033559, R56AG026622-01A1, and RC4AG039036), with additional support from the E. Morris Cox Fund at the University of California, Berkeley. We gratefully acknowledge comments by James P. Smith and other participants of the NBER Economics of Aging Conference, Carefree, AZ, May 2015. We thank Patricia St. Clair for invaluable help with the data and Ingrid Hagele and Katrin Poschen for excellent research assistance. For acknowledgments, sources of research support, and disclosure of the author's or authors' material financial relationships, if any, please see http://www.nber.org/chapters/c13637.ack.

This work extends an existing literature that compares survey self-reports and biomarker-based measures of disease prevalence. These papers focus on diabetes (Goldman et al. 2003; Baker, Stabile, and Deri 2004; Smith 2007; Barcellos, Goldman, and Smith 2012; Chatterji, Joo, and Lahiri 2012) and/or hypertension (Goldman et al. 2003; Johnston, Propper, and Shields 2009; Barcellos, Goldman, and Smith 2012; Chatterji, Joo, and Lahiri 2012). These are all diseases for which biomarkers can be obtained relatively easily in community surveys such as the HRS.[1] Data linkage provides another opportunity to verify survey self-reports. Two recent studies, Wolinsky et al. (2014) and Yasaitis, Berkman, and Chandra (2015), compare survey self-reports of different conditions with diagnoses documented in Medicare claims data that have been linked to the HRS data. Sakshaug, Weir, and Nicholas (2014) are the first to compare measures from all three data sources: self-reports, biomarkers, and claims data. They document large differences in diabetes prevalence between HRS self-reports and linked Medicare claims and show that self-reported diabetes aligns more closely with the biomarker data. Taking the biomarker data as a "gold standard" they conclude that diabetes prevalence in the Medicare claims data is too high. The present chapter takes a closer look at the three different measures of diabetes in the HRS, biomarker, and linked Medicare claims data. In particular, our analysis takes the perspective that all three measures may suffer from measurement error.

Substantively, the results from prior literature show that survey respondents tend to underreport the prevalence of diabetes and hypertension compared to "objective" measures from biomarkers. There are socioeconomic status (SES) gradients both in prevalence itself and in the measurement error contained in self-reports, but they are not necessarily the same.

Goldman et al. (2003) find in data from Taiwan that survey self-reports vastly underestimate the prevalence of hypertension, but yield a reasonable accurate estimate of diabetes prevalence. The accuracy of self-reports is predicted by age, education, time of the most recent health exam, and cognitive function.

For the United States, Smith (2007) documents predictors of diabetes prevalence and undiagnosed diabetes using data from three National Health and Nutrition Examination Survey (NHANES) waves. He finds that diabetes prevalence is predicted primarily by excessive weight and obesity. Inheritance of diabetes through parents is also important. These forces were only partially offset by improvements in the education of the population over time. Further, Smith shows that about one in five male diabetics were

[1] Other surveys used in related studies include the National Health and Nutrition Examination Survey (NHANES) in Smith (2007) and Barcellos, Goldman, and Smith (2012); the Health Survey for England (HSE) in Johnston, Propper, and Shields (2009); and the Canadian National Population Health Survey (NPHS) in Baker, Stabile, and Deri (2004). A related study that uses the HRS is Chatterji, Joo, and Lahiri (2012).

undiagnosed in the 1999–2002 NHANES waves. While race and ethnic differentials in undiagnosed diabetes were eliminated over the last twenty-five years, the disparities became larger across other measures of disadvantage such as education. Undiagnosed diabetes is a particularly severe problem among the obese, a group at much higher risk of diabetes onset. Also for the United States and with NHANES data, Barcellos, Goldman, and Smith (2012) study undiagnosed diabetes among Mexican immigrants. The striking finding is that these immigrants might be much less healthy than previously thought because diseases remain undiagnosed at a much higher rate than among other groups of the US population. With respect to diabetes, Barcellos et al. document that about half of recent immigrants with the disease remain undiagnosed.

An important issue is whether the measurement error contained in survey self-reports is related to socioeconomic status (SES). The findings in Smith (2007) and Barcellos, Goldman, and Smith (2012) suggest that this is indeed the case for diabetes in the United States. Similar SES gradients in undiagnosed hypertension have been documented by Johnston, Propper, and Shields (2009) for England.

Curiously, when comparing self-reports and insurance claims data, Wolinsky et al. (2014), and Yasaitis, Berkman, and Chandra (2015) document that the measurement error in survey self-reports may also go the other way: Wolinsky et al. (2014) document over- as well as underreporting of different health conditions and health care use in survey data as compared to claims. The authors find an SES gradient with respect to wealth in the accuracy of the self-reports. Yasaitis, Berkman, and Chandra (2015) focus on acute myocardial infarctions (AMI) and find that less than half of those who reported a heart attack in their HRS sample had evidence of acute cardiovascular hospitalizations in the Medicare claims data. Further, they did not find associations between demographic characteristics and the frequency with which self-reported AMI was verified by Medicare claims.

Not only survey self-reports, but also measures of disease prevalence constructed from claims data may be subject to measurement error. Sakshaug, Weir, and Nicholas (2014) find that roughly 8 percent of HRS respondents in 2006 do not report having been diagnosed with diabetes but are identified as diabetics based on the linked Medicare claims. Among these cases, almost 64 percent do not have diabetes according to the available biomarker information, suggesting that the procedure of identifying diabetes cases in the claims data may lead to false positives.

The literature thus suggests that neither self-reports nor claims data may deliver reliable measures of health conditions and that both measures may be subject to Type I and Type II errors. The possibility of measurement error in the biomarker data, however, has received less attention. The present chapter further explores this issue. Furthermore, the question of whether the different errors are predicted by SES is still open and the present chapter presents some additional evidence. A potentially important consideration, which we will not address in this chapter, is to what extent selectivity in linked samples—due to incomplete consent of survey respondents to either biomarker measurement or to claims data linkage—contributes to observed SES gradients in the various measures of disease prevalence and the associated measurement errors.

The remainder of the chapter is structured as follows: We discuss the data used in section 7.2. Section 7.3 contains the results, and section 7.4 concludes with a summary of our findings and a discussion of avenues for future research.

7.2 Data

In the analysis presented in the chapter, we use three linked data sets: biomarker data on diabetes prevalence are taken from the HRS (2006 and 2008), self-reports of diabetes are taken from the RAND HRS data set (where SES covariates are readily defined), and claims-based information on diabetes prevalence is taken from Medicare claims that are linked to HRS respondents. The latter data are provided by the Medicare Research Information Center (MedRic). Column (1) of table 7.1 displays the number of observations in each of these data sets. Column (2) shows the number of observations with nonmissing diabetes indicators. Comparing these two columns shows that the rates of missing items (due to nonresponse or other reasons) are low.

Columns (3), (4), and (5) contain diabetes prevalence in each of the data sets. In the biomarker data, we use glycated hemoglobin (HbA1c) levels with 6.5 percent as a threshold (the NHANES equivalent value);[2] in the RAND HRS data, we use the self-reported diagnosed diabetes; in the claims data, we use an "ever had" diabetes claims indicator as provided by MedRIC. The latter is based on the procedure used to identify diabetes in other commonly used Medicare claims data by the Chronic Conditions Data Ware-

[2] The HbA1c level captures chronic hyperglycemia and has traditionally been used to monitor diabetes treatment (e.g., Bonora and Tuomilehto 2011). In this respect, the American Diabetes Association recommends that HbA1c levels are regularly checked for diabetes patients and patients should try to reach specific HbA1c target levels (usually < 7 percent, but for some patients lower targets, such as <6.5 percent, may be appropriate) as lower HbA1c levels are associated with lower risk of diabetes-related complications. The use of HbA1c screenings as a diagnostic tool for diabetes has only started recently, after extensive research had demonstrated its value for identifying undiagnosed diabetes (e.g., Rohlfing et al.2000; Bennett, Guo, and Dharmage 2007) although other authors conclude that it is not a reliable measure to detect diabetes (Reynolds, Smellie, and Twomey 2006). In the United States, the American Diabetes Association started to recommend HbA1c screening as a test for diabetes in 2010 (American Diabetes Association 2010). HbA1c levels below 5.7 percent are considered normal, levels 5.7–6.5 percent are considered as prediabetes, and levels above 6.5 percent indicate diabetes (American Diabetes Association 2015).

Table 7.1 Prevalence of diabetes—Comparisons across measures

	(1)	(2)	(3)	(4)	(5)
			Diabetes indicators		
	N	*N*	HbA1c> = 6.5%	Self-reported (%)	Claims (%)
2006					
Biomarkers	6,735	6,517	12.38	n/a	n/a
HRS	18,469	18,435	NA	19.60	n/a
MedRIC (HRS-claims)	11,323	11,323	n/a	n/a	25.34
2008					
Biomarkers (total)	6,329	6,256	15.04	n/a	n/a
Biomarkers (biosafe lab)		4,347	14.49	n/a	n/a
Biomarkers (flex lab)		1,909	16.29	n/a	n/a
HRS	17,217	17,185	n/a	21.57	n/a
MedRIC (HRS-claims)	10,597	10,597	n/a	n/a	29.86
Linked sample					
All individuals					
2006	4,118	3,956	16.66	22.27	24.32
2008	3,904	3,853	18.82	24.27	28.13
Excluding individuals in HMOs					
2006		2,517	16.01	21.97	26.36
2008		2,370	18.99	22.70	30.51

Notes: Column (1) displays overall numbers of observations in each data set. Column (2) limits to those observations with information on diabetes. Diabetes indicators in the claims data are based on the Chronic Conditions Data Warehouse (CCW) definitions; those based on HbA1c biomarkers use NHANES equivalent definitions.

house (CCW).[3] In the 2006 biomarker data, for instance, 12.38 percent of respondents have diabetes according to their HbA1c level, while 19.6 percent of the respondents from 2006 HRS report ever having been diagnosed with diabetes in the HRS. In the linked MedRIC claims data, 25.34 percent of individuals are identified as diabetics. As these samples contain different individuals that vary in age, for example, the rates are, however, not directly comparable across data sources.

As biomarker data are collected in the HRS only every second wave, the biomarker samples are substantially smaller than the full HRS. In addition, not all respondents consent to biomarker measurement and/or claims data

[3] However, in the MedRIC claims data, the diabetes flag is coded as either 0 or 1 with no missing values, while in the other CMS Medicare data, it is coded in 4 levels: (a) incomplete claims coverage for the reference period and diagnosis not found; (b) incomplete claims coverage for the reference period and diagnosis found; (c) complete claims coverage for the reference period and diagnosis not found; and (d) complete claims coverage for the reference period and diagnosis found. In the MedRIC claims data, we do not know how the cases with incomplete claims were coded.

linkage, which results in a further loss of survey cases. However, consent rates in the HRS are generally high.[4] After merging the three data sets, we have 4,118 observations for year 2006 and 3,904 observations for year 2008. Excluding cases with missing information on diabetes-related variables, the linked data contain 3,956 individuals in 2006 and 3,853 individuals in 2008. This includes individuals with different types of Medicare coverage, in particular, individuals who are in traditional Medicare (in a fee-for-service [FFS] plan) and individuals who are in a Health Maintenance Organization (HMO) through Medicare Advantage. For individuals in HMOs we do not observe all relevant claims and we thus conduct most analyses excluding this group of individuals.[5] Excluding HMO individuals and focusing on nonmissing diabetes-related variables, we have 2,517 observations for 2006 and 2,370 observations for 2008.

Columns (3), (4), and (5) of the bottom panel of table 7.1 display diabetes prevalence rates for the linked samples. Even in the linked sample diabetes prevalence is lowest according to the biomarker data and highest in the claims data. This pattern is similar in both years and aligns with the findings of Sakshaug, Weir, and Nicholas (2014), who only analyzed the 2006 HRS data.[6] We explore these patterns in more detail in the next section.

7.3 Results

We first consider diabetes prevalence by gender and by educational levels (table 7.2). For education, we use the five education categories in the RAND HRS data: less than high school; GED; high school graduates; some college; college and above. Results in the lower panels of table 7.2 show educational gradients in diabetes based on all three diabetes indicators. For both genders and in all education groups in both years, diabetes prevalence is lowest according to the HbA1c and highest in the claims data.

Table 7.3 shows two-way within-respondent comparisons of the different measures for the years 2006 and 2008. For all measures and in both years, the concordance across different measures is quite high. The first panel,

[4] According to the HRS biomarker documentation, in 2006 the consent rate for obtaining dried blood spots from which the HbA1c measure is extracted was 83 percent and the completion rate, conditional on consent, was 97 percent. The overall completion rate was 81 percent. In 2008, the overall completion rate was 87 percent. According to the MedRIC documentation, over 80 percent of all HRS respondents who are eligible for Medicare provided their identification numbers so that claims data could be linked.

[5] In the analyses that do not include information in the claims data, this restriction is not necessary. The results are almost identical when individuals in HMOs are included and thus not discussed further. They are available upon request.

[6] Sakshaug et al. also exclude individuals in HMOs. In addition, they restrict their analysis to individuals older than sixty-five who are not veterans. Although we do not implement these additional restrictions, our results are almost identical. While Sakshaug et al. report that 27.3 percent in their linked sample have diabetes according to the claims data in 2006, in our sample definition 26.4 percent are identified as diabetics in the claims.

Table 7.2 Diabetes by gender and education

			Diabetes indicators		
	N	(%)	HbA1c> = 6.5%	Self-reported (%)	Claims (%)
		Gender			
2006					
Male	1,079	42.89	14.92	23.63	27.53
Female	1,437	57.11	12.38	20.72	25.45
	Chi-square test		3.4204	3.0445	1.3660
	P-value		0.0640	0.0810	0.2430
2008					
Male	997	42.41	18.65	27.08	33.73
Female	1,354	57.59	14.39	19.46	28.12
	Chi-square test		7.7434	19.2020	8.5989
	P-value		0.0050	0.0000	0.0030
		Education			
2006					
Less than high school	515	20.47	18.06	29.13	35.53
GED	124	4.93	19.35	23.39	28.23
High school graduate	840	33.39	12.59	20.78	26.48
Some college	516	20.51	9.90	18.83	22.33
College and above	521	20.70	12.48	19.58	20.54
	Chi-square test		19.612	20.912	35.982
	P-value		0.001	0.000	0.000
2008					
Less than high school	520	22.12	23.53	32.26	41.18
GED	108	4.59	17.59	30.56	41.67
High school graduate	777	33.05	14.32	21.74	29.67
Some college	455	19.35	16.96	18.26	25.43
College and above	491	20.89	10.37	16.46	22.56
	Chi-square test		35.552	47.704	55.167
	P-value		0.000	0.000	0.000

Notes: Excluding individuals who have Medicare coverage through an HMO.

for example, compares diabetes according to the biomarker data and self-reports in 2006 and 2008. In 2006, roughly 10 percent of individuals report having been diagnosed with diabetes and have an HbA1c level higher than 6.5 percent. Another 75 percent of individuals report not having been diagnosed with diabetes and have an HbA1c level lower than 6.5 percent. For 85 percent of cases, self-reports and biomarker data thus align. The respective results for 2008 are almost identical. In both years, 15 percent of respondents have inconsistent results according to the two diabetes indicators. In 2006, 3.26 percent (4.05 percent in 2008) of respondents have HbA1c levels higher than 6.5 percent but do not report diabetes (which may reflect cases of undiagnosed diabetes) while 11.76 percent (10.55 percent in 2008) of the individuals have HbA1c levels lower than 6.5 percent but report having

Table 7.3 **Comparison of measures of diabetes**

	(%)	(%)
	HRS self-reported diabetes and biomarker	
2006 (N = 2,517)	HRS self-reported diabetes	
Biomarker	Yes	No
HbA1c level (%) > = 6.5	10.21	3.26
HbA1c level (%) < 6.5	11.76	74.77
2008 (N = 2,370)		
HbA1c level (%) > = 6.5	12.15	4.05
HbA1c level (%) < 6.5	10.55	73.25
	Diabetes according to claims and biomarker	
2006 (N = 2,517)	Diabetes according to claims	
Biomarker	Yes	No
HbA1c level (%) > = 6.5	9.57	3.89
HbA1c level (%) < 6.5	16.77%	69.77
2008 (N = 2,370)		
HbA1c level (%) > = 6.5	12.45	3.76
HbA1c level (%) < 6.5	18.06	65.74
	HRS self-reported diabetes and according to claims	
2006 (N = 2,517)	Diabetes according to claims	
Self-reports	Yes	No
HRS self-reported diabetes: Yes	18.87	3.10
HRS self-reported diabetes: No	7.47	70.56
2008 (N = 2,370)		
HRS self-reported diabetes: Yes	20.93	1.77
HRS self-reported diabetes: No	9.58	67.72

Notes: Excluding individuals who have Medicare coverage through an HMO.

been diagnosed with diabetes. The latter cases may reflect overdiagnoses or diabetes cases that are successfully treated. We explore the possibility of under- and overdiagnosis in the self-reports in more detail below (in tables 7.4 and 7.5).

In the middle panel of table 7.3, we compare diabetes according to biomarkers and claims data. In this comparison, roughly 21 percent of respondents have inconsistent results; 3.89 percent of the sample in 2006 and 3.76 percent in 2008 have no diabetes claims, yet have HbA1c levels that exceed 6.5 percent; and 16.77 percent in 2006 and 18.06 percent in 2008 have diabetes claims, but their HbA1c level is below 6.5 percent.

The bottom panel of table 7.3 compares HRS self-reported diabetes with diabetes according to the claims data. The discrepancies are even smaller than when comparing the other measures: 3.1 percent of the sample in 2006 and 1.77 percent in 2008 report ever having been diagnosed with diabetes but have no diabetes claims, while 7.47 percent of the respondents in 2006 and 9.58 percent in 2008 report not having been diagnosed with diabetes but are identified as diabetic in the claims data. The latter findings are again very similar to Sakshaug, Weir, and Nicholas (2014) who report that in 2006 7.7 percent of individuals have diabetes according to the claims data but do not report having been diagnosed with diabetes.

In tables 7.4 and 7.5, we try to reconcile the discrepancies that arise when comparing the self-reports and the biomarker data. Table 7.4 focuses on the possibly "overdiagnosed" cases, while table 7.5 focuses on the possibly "undiagnosed" cases. Table 7.4 displays self-reported medical treatment for individuals who report that they have been diagnosed with diabetes but do not have diabetes according to the biomarker data. In both years, a large fraction among these individuals report taking swallowed medication (almost 74 percent in 2006 and 69 percent in 2008). Furthermore, between 13 and 14 percent report being treated with insulin. Combining the two treatments, 81 percent in 2006 and 77 percent in 2008 report being treated for diabetes. A majority of the differences between self-reports and biomarker data in diabetes may thus stem from successfully treated diabetes cases rather than overreporting in the self-reported data. This is also plausible, as the American Diabetes Association (2015), for example, suggests that providers may recommend patients to target HbA1c levels below 6.5 percent, as this lowers the risk of diabetes-related complications.

Table 7.5 focuses on individuals with high HbA1c levels who do not report having diabetes. There are two main explanations for why individuals do not report diabetes while their HbA1c levels are above 6.5 percent. First, they may have been diagnosed with diabetes but they simply forget to—or do not want to—mention it during the HRS interview. Second, they may not know that they have diabetes. While for individuals who report not having been diagnosed with diabetes there is no information on treatment in the HRS survey, we can look at the claims data to investigate whether these individuals receive treatment for diabetes and have taken diabetes screenings. Table

Table 7.4 **Reconciliation HRS self-reports and biomarker information–Medical treatment among seemingly false positive self-reports**

	Swallowed medication			Insulin			Either of the two treatments		
	Yes (%)	No (%)	Missing (%)	Yes (%)	No (%)	Missing (%)	Yes (%)	No (%)	Missing (%)
2006 ($N = 296$)	73.65	25.68	0.68	14.53	84.80	0.68	81.42	17.91	0.68
2008 ($N = 250$)	68.80	30.80	0.40	13.20	86.40	0.40	76.80	22.80	0.40

Notes: Excluding individuals who have Medicare coverage through an HMO.

Table 7.5 **Reconciliation HRS self-reports and biomarker information—Claims and diabetes screening among seemingly false negative self-reports**

2006	*N* = 82 (%)		*N* = 22 (%)	*N* = 21 (enrolled in Medicare for at least 2 years before earliest diabetes diagnosis) (%)	
Diabetes claims: Yes	26.83	Percent ever had glucose test before earliest indication of diabetes	45.45	47.62	
		Percent ever had HbA1c test before earliest indication of diabetes	68.18	61.90	
		Percent ever had screening test before earliest indication of diabetes	81.82	76.19	
Diabetes claims: No	73.17	Percent ever had glucose test	23.33	Percent had glucose test in recent two yrs.	11.67
		Percent ever had HbA1c test	15.00	Percent had HbA1c test in recent two yrs.	8.33
		Percent ever had screening test before the HRS 2006 interview	31.67	Percent ever had screening test in recent two yrs.	18.33
2008	*N* = 96		*N* = 23	*N* = 22 (enrolled in Medicare for at least two years before earliest diabetes diagnosis)	
Diabetes claims: Yes	23.96	Percent ever had glucose test before earliest indication of diabetes	43.48	40.91	
		Percent ever had HbA1c test before earliest indication of diabetes	78.26	68.18	
		Percent ever had screening test before earliest indication of diabetes	82.61	72.73%	
Diabetes claims: No	76.04	Percent ever had glucose test	31.51	Percent had glucose test in recent two yrs.	6.85
		Percent ever had HbA1c test	23.29	Percent had HbA1c test in recent two yrs.	6.85
		Percent ever had screening test before the HRS 2008 interview	43.84	Percent ever had screening test in recent two yrs.	12.33

Notes: Excluding individuals who have Medicare coverage through an HMO.

7.5 shows that 26.8 percent of "undiagnosed" cases in 2006 and almost 24 percent in 2008 are identified as diabetics in the claims data. As this suggests that individuals receive treatment for diabetes, the fraction of truly undiagnosed cases reduces from 3.26 percent to 2.4 percent in 2006 and from 4.05 percent to 3.1 percent in 2008. Diabetes screenings are identified in the claims based on CPT-4 and ICD-9 diagnosis codes. As the list shown below indicates, we identify two types of tests: a glucose test and a glycated hemoglobin (HbA1c) test. Furthermore, there is a general code for diabetes screening: 82947 Assay Body Fluid Glucose; 82950 Glucose Test; 82951 Glucose Tolerance Test (GTT); 83036 Glycated Hemoglobin (HbA1c) Test; and V77.1 Screen for diabetes mellitus.

Based on this information, we construct three indicators: (a) whether an individual has taken a glucose test, (b) whether an individual has taken an HbA1c test, and (c) whether an individual has taken a glucose or HbA1c test. For individuals who have high HbA1c levels but do not report having been diagnosed with diabetes and are not identified as diabetic in the claims data, we investigate whether they have taken a screening test before the HRS survey date. We further check whether this group of people has taken a screening test in the two years before their HRS interview. For individuals who have high HbA1c levels and diabetes claims but do not report diabetes, we show the fraction of individuals who took the different screening tests before the onset of diabetes in their claims records.

The results in table 7.5 indicate that 32 percent of individuals with high HbA1c levels but no self-reported diabetes and no claims in 2006 and 44 percent among these individuals in 2008 have taken at least one diabetes screening test before their HRS interview. When restricting it to a two-year time horizon before the HRS interview, only 18 percent in 2006 and 12 percent in 2008 have taken a test. This compares to 82–83 percent among individuals with diabetes claims. This suggests that a large fraction of individuals with high biomarker data but no diabetes according to self-reports or claims data truly have undiagnosed diabetes.

Next, we study how having undiagnosed diabetes varies by gender. Table 7.6 displays within comparisons of the biomarker and self-reported diabetes measures by gender and year. Furthermore, it investigates for potential cases of undiagnosed diabetes (high HbA1c but no diabetes according to self-reports) whether individuals are identified as diabetics based on their Medicare claims. The results are very similar across gender and years. Self-reported diabetes and diabetes in the biomarker data align for roughly 85 percent of individuals. The fraction of potentially false positives (or overdiagnoses) in the self-reported data is between 10 and 11 percent; the fraction of potentially false negatives (or undiagnosed cases) is 3–4 percent. The fraction who have high HbA1c levels, no self-reports, and also do not have diabetes according to their Medicare claims varies across genders and years. However, these differences are not statistically significantly different

Table 7.6 **Comparison of self-reports, biomarker data, and claims by gender**

	HRS self-reported diabetes: Male		HRS self-reported diabetes: Female	
	Yes (%)	No (%)	Yes (%)	No (%)
2006	*N* = 1,097		*N* = 1,438	
HbA1c level (%) > = 6.5	11.49	3.19	9.11	3.27
HbA1c level (%) < 6.5	11.76	71.92	11.61	76.01
	3.19 (obs. = 35)	3.27 (obs. = 47)		
	Diabetes according to claims: Yes	34.29	Diabetes according to claims: Yes	21.28
	Diabetes according to claims: No	65.71	Diabetes according to claims: No	78.72
2008	*N* = 1,008		*N* = 1,362	
HbA1c level (%) > = 6.5	15.28	3.37	9.84	4.55
HbA1c level (%) < 6.5	11.81	69.51	9.62	75.99
	3.37 (obs. = 34)	4.55 (obs. = 62)		
	Diabetes according to claims: Yes	20.59	Diabetes according to claims: Yes	25.81
	Diabetes according to claims: No	79.41	Diabetes according to claims: No	74.19

Notes: Excluding individuals who have Medicare coverage through an HMO.

from zero.[7] Overall, we do not find evidence of gender differences in undiagnosed diabetes.

An important issue in the literature on diabetes prevalence is its gradient in SES. We study whether SES is a predictor of diabetes prevalence as well as prevalence of undiagnosed diabetes among all diabetics. We define the latter two measures based on different combinations of the three available measures. Tables 7.7 and 7.8 present summary statistics of different demographic and socioeconomic variables as measured in the HRS data for 2006 and 2008, respectively.

The means of the variables are presented for the entire linked samples (excluding individuals in HMOs) and separately for individuals with and without diabetes according to the three different measures. The last three columns of tables 7.7 and 7.8 display means for the groups of individuals who potentially have undiagnosed diabetes according to three different definitions that we discuss further below.

For both years and across all three measures of diabetes, tables 7.7 and 7.8 suggest that race, ethnicity, education, earnings, wealth, and self-assessed health status are associated with diabetes prevalence. Among individuals with diabetes, a lower share is white and a higher share is Hispanic. In addition, diabetics have on average lower education, lower income, and lower wealth than nondiabetics and more of them rate their health as fair or poor. Furthermore, as one would expect, individuals with diabetes have a higher body mass index (BMI) on average and fewer among them report doing vigorous exercise.

While tables 7.7 and 7.8 analyze each diabetes measure separately, one could also combine the three available measures in different ways. The following analysis is a first attempt in that direction. We start with counting everyone as diabetic who either reports having been diagnosed with diabetes, or has an HbA1c that exceeds 6.5 percent, or has diabetes in the claims data. However, we also rely on self-reports and biomarker data alone to facilitate comparison of our results to the earlier literature that had no matched claims data. Table 7.9 provides means of the demographic, socioeconomic, and health- and health insurance-related variables for these two different definitions of diabetes and both years of data. To simplify comparisons, table 7.9 also displays means for individuals that are only identified as diabetics in the claims data. The results suggest that the latter group is slightly older, more likely to be white and female, has higher wealth, better self-rated health, and slightly lower BMI compared to individuals who either report having been diagnosed with diabetes or have high HbA1c levels. In addition, they naturally have a lower HbA1c level on average compared to the other groups of diabetics. While this group may indeed include certain individuals who are falsely classified as diabetics as Sakshaug, Weir, and Nicholas

[7] Results available upon request.

Table 7.7 **Summary statistics 2006**

Variable	Whole linked sample	HbA1c > = 6.5%	HbA1c < 6.5%	SR diabetes: Yes	SR diabetes: No	Diabetes claims: Yes	Diabetes claims: No	Undiagnosed diabetes 1: High HbA1c no self-reports	Undiagnosed diabetes 2: High HbA1c, no self-reports, no claims	Undiagnosed diabetes:3: High HbA1c, no self-reports, no claims, no screening in past two years
Age	72.80	71.98	72.92	71.93	73.04	73.82	72.43	73.21	72.67	72.08
Age > = 65	0.88	0.86	0.89	0.86	0.89	0.93	0.87	0.91	0.90	0.90
Race (white)	0.88	0.79	0.89	0.82	0.89	0.83	0.90	0.74	0.80	0.78
Hispanic	0.05	0.10	0.04	0.09	0.04	0.09	0.04	0.05	0.05	0.06
Female	0.57	0.53	0.58	0.54	0.58	0.55	0.58	0.57	0.62	0.61
Married	0.63	0.60	0.64	0.62	0.64	0.60	0.65	0.59	0.57	0.57
Education: High school and above	0.80	0.73	0.81	0.73	0.81	0.72	0.82	0.76	0.78	0.78
Indiv. earnings	5,449.77	4,301.37	5,628.52	3,436.00	6,016.79	2,706.23	6,430.88	5,067.02	6,658.27	7,882.57
HH income	57,303.95	44,182.31	59,346.30	45,022.89	60,761.90	42,544.32	62,582.07	41,909.11	42,702.33	45,441.74
HH income (median)	36,380.00	31,988.00	37,381.00	31,060.00	38,400.00	31,164.00	38,376.00	29,533.27	27,404.00	30,651.00
HH wealth	577,488.4	413,044.5	603,083.7	429,214.0	619,237.8	428,920.2	630,617.2	319,767.9	317,605.80	337,720.0
HH wealth (median)	264,400.0	155,000.0	284,152.0	177,000.0	299,530.0	177,000.0	300,000.0	173,763.0	195,040.0	195,080.0
Covered by EGHP	0.46	0.41	0.47	0.43	0.47	0.43	0.47	0.39	0.37	0.41
Poor/fair general health status	0.29	0.41	0.27	0.45	0.25	0.42	0.24	0.26	0.23	0.22
BMI	27.71	30.18	27.32	30.44	26.93	29.36	27.11	28.92	29.06	29.58
Total cognition	22.03	21.26	22.15	21.31	22.23	21.05	22.41	20.96	21.69	21.66
Ever smoker	0.58	0.58	0.58	0.59	0.57	0.59	0.57	0.50	0.53	0.53
Current smoker	0.11	0.10	0.11	0.10	0.12	0.10	0.12	0.09	0.07	0.08
Vigorous exercise	0.34	0.24	0.35	0.26	0.36	0.27	0.36	0.30	0.32	0.31
Part A enrollment	0.92	0.94	0.92	0.93	0.91	0.99	0.89	0.98	0.97	0.96
Part B enrollment	0.86	0.89	0.86	0.89	0.85	0.96	0.83	0.91	0.91	0.90
Drug coverage (part D or other sources)	0.79	0.79	0.79	0.81	0.78	0.85	0.76	0.80	0.85	0.82
Medicaid	0.07	0.12	0.06	0.11	0.06	0.11	0.06	0.10	0.10	0.06
Observations	2,517	339	2,178	553	1,964	663	1,854	82	60	49

Notes: Excluding individuals who have Medicare coverage through an HMO. Means of respective variables, unless indicated otherwise. EGHP indicates employer-sponsored health insurance.

Table 7.8 **Summary statistics 2008**

Variable	Whole linked sample	HbA1c > = 6.5%	HbA1c < 6.5%	SR diabetes: Yes	SR diabetes: No	Diabetes claims: Yes	Diabetes claims: No	Undiagnosed diabetes 1: High HbA1c no self-reports	Undiagnosed diabetes 2: High HbA1c, no self-reports, no claims	Undiagnosed diabetes:3: High HbA1c, no self-reports, no claims, no screening in past two years
Age	73.88	72.95	74.06	72.48	74.29	74.16	73.76	74.31	73.41	73.33
Age > = 65	0.95	0.93	0.95	0.91	0.96	0.93	0.96	0.97	0.96	0.95
Race (white)	0.87	0.77	0.89	0.80	0.89	0.82	0.89	0.78	0.77	0.78
Hispanic	0.07	0.11	0.06	0.11	0.05	0.11	0.05	0.07	0.04	0.05
Female	0.57	0.51	0.59	0.49	0.60	0.53	0.59	0.65	0.63	0.63
Married	0.56	0.53	0.57	0.53	0.57	0.51	0.58	0.54	0.59	0.58
Education: High school and above	0.78	0.68	0.80	0.68	0.81	0.70	0.81	0.77	0.78	0.75
Indiv. earnings	4,370.63	4,172.96	4,408.85	3,652.10	4,581.64	2,356.04	5,255.00	6,931.84	8,850.69	8,884.38
HH income	53,895.97	43,758.34	55,856.11	43,815.66	56,856.23	41,008.84	59,553.16	51,761.80	58,736.68	57,430.32
HH income (median)	35,009.00	30,300.00	36,000.00	28,804.00	37,936.00	28,804.00	37,936.00	32,052.00	41,235.00	41,197.50
HH wealth	577,099.4	401,678.5	611,017.6	356,055.1	642,013.1	393,180.8	657,835.9	395,058.30	439,706.5	428,768.7
HH wealth (median)	232,750.0	123,750.0	262,500.0	139,000.0	288,500.0	139,000.0	288,500.0	169,000.0	232,000.0	234,000.0
Covered by EGHP	0.43	0.39	0.43	0.42	0.43	0.41	0.44	0.42	0.48	0.45
Poor/fair general health status	0.32	0.49	0.28	0.48	0.27	0.46	0.25	0.33	0.32	0.28
BMI	27.87	30.64	27.34	30.60	27.07	29.84	27.01	29.03	29.53	29.56
Cognition	21.80	20.84	21.98	20.96	22.04	20.74	22.26	21.52	22.33	22.13
Ever smoker	0.58	0.61	0.58	0.64	0.57	0.62	0.57	0.51	0.47	0.44
Current smoker	0.11	0.11	0.11	0.11	0.11	0.10	0.12	0.10	0.11	0.09
Vigorous exercise	0.33	0.23	0.35	0.27	0.35	0.26	0.36	0.24	0.27	0.28
Part A enrollment	0.99	0.98	0.99	0.98	0.99	0.99	0.99	0.99	1.00	1.00
Part B enrollment	0.93	0.93	0.93	0.92	0.93	0.96	0.92	0.92	0.92	0.92
Drug coverage (part D or other sources)	0.88	0.90	0.88	0.91	0.88	0.91	0.87	0.90	0.93	0.92
Medicaid	0.10	0.17	0.08	0.18	0.07	0.17	0.06	0.09	0.06	0.06
Observations	2,370	384	1,986	538	1,832	723	1,647	97	73	64

Notes: Excluding individuals who have Medicare coverage through an HMO. Means of respective variables unless indicated otherwise.

Table 7.9 **Summary statistics—Cases with diabetes according to different measures**

	2006			2008		
Variable	Self-reports or biomarker	Self-reports or biomarker or claims	Claims only	Self-reports or biomarker	Self-reports or biomarker or claims	Claims only
Age	72.09	73.02	76.55	72.76	73.80	77.09
Age > = 65	0.87	0.89	0.98	0.92	0.93	0.96
Race (white)	0.81	0.82	0.84	0.80	0.81	0.87
Hispanic	0.09	0.08	0.07	0.11	0.10	0.08
Female	0.54	0.56	0.61	0.52	0.54	0.61
Married	0.62	0.61	0.57	0.53	0.53	0.51
Education: High school and above	0.73	0.74	0.75	0.70	0.71	0.76
HH wealth	415,080.8	420,042.10	439,020.44	362,013.10	399,387.3	513,974.5
HH wealth (median)	162,000.0	175,000.0	234,550.0	125,000.0	155,000.0	228,408.7
Covered by EGHP	0.43	0.43	0.46	0.42	0.42	0.42
Poor/fair general health status	0.42	0.40	0.33	0.46	0.44	0.39
BMI	30.25	29.72	27.69	30.36	29.79	28.01
Cognition	21.26	21.17	20.86	21.05	20.99	20.82
Ever smoker	0.58	0.59	0.60	0.62	0.60	0.56
Current smoker	0.10	0.10	0.10	0.11	0.10	0.10
Vigorous exercise	0.27	0.27	0.29	0.26	0.26	0.27
Part A enrollment	0.93	0.95	0.99	0.98	0.99	1.00
Part B enrollment	0.89	0.91	0.96	0.92	0.94	0.98
Drug coverage (part D or other sources)	0.81	0.82	0.85	0.91	0.91	0.91
Medicaid	0.11	0.11	0.08	0.16	0.15	0.12
Observations	635	801	166	635	839	204

Notes: Excluding individuals who have Medicare coverage through an HMO. Means of respective variables unless indicated otherwise.

(2014) suggest, it may also include individuals whose diabetes is under control through treatment and who thus do not report having diabetes.

The results presented in tables 7.10 and 7.11 explore predictors of diabetes in multivariate regressions. In table 7.10, all individuals who have diabetes according to either of the three measures are classified as diabetics. In table 7.11, diabetes is defined according to self-reports only, using biomarker data as a predictor for self-reported diabetes. Both tables include average marginal effects after probit estimation for the years 2006 (columns [1]–[3]) and 2008 (columns [4]–[6]). For each year, the first column displays results for demographic and socioeconomic variables only, the second column adds health indicators and health behavior as explanatory variables, and the third column adds information on an individual's insurance status.

The results in table 7.10 confirm many of the findings from the descriptive analyses in tables 7.7 and 7.8 and are in line with the earlier literature studying predictors of diabetes. Individuals who are white are less likely to have diabetes, and those who are Hispanic are more likely to have diabetes. In addition there are gradients in education and wealth, although the former is no longer significantly different from zero when health indicators and behaviors are included. Individuals who rate their health as fair or poor have a 10 to 13 percentage points higher probability of having diabetes compared to individuals who rate their health as better. In addition, diabetes risk is positively related to BMI and negatively to doing vigorous exercise. These associations do not change when insurance status is included. The results in table 7.11 indicate that the HbA1c level is a very significant predictor of self-reported diabetes, even when controlling for all the other demographic, socioeconomic, health- and health insurance-related variables. Many of the latter variables do not significantly predict self-reported diabetes when HbA1c is included (e.g., race, ethnicity), mainly self-reported general health and BMI remain robust and strong predictors above and beyond the HbA1c level.[8]

The descriptive analysis above already suggested that measures of undiagnosed diabetes depend on the data sources and how they are combined. In the final set of analyses we therefore study alternative definitions of undiagnosed diabetes that combine information from all three sources of prevalence data in different ways:

1. Undiagnosed1: HbA1c level > = 6.5 percent, but no self-reported diabetes.
2. Undiagnosed2: HbA1c level > = 6.5 percent, but no self-reported diabetes and no diabetes in claims data.

[8] For the results presented in tables 7.10 and 7.11, we additionally explored changes when using categories of BMI instead of the linear value, when restricting the sample to individuals age sixty-five and older, and when replacing household wealth with household income. The results were remarkably similar and are thus not shown here. They are, however, available upon request.

Table 7.10 Probability of diabetes (according to any of the three measures)—Marginal effects after probit estimation

	2006			2008		
	SES	+health indicators	+insurance status	SES	+health indicators	+insurance status
Age	0.002*	0.003**	0.003*	0.000	0.003**	0.003*
	[0.0012]	[0.0015]	[0.0016]	[0.0013]	[0.0016]	[0.0016]
White	–0.148***	–0.081**	–0.089***	–0.140***	–0.094***	–0.101***
	[0.0306]	[0.0327]	[0.0335]	[0.0314]	[0.0324]	[0.0330]
Hispanic	0.162***	0.128***	0.130**	0.136***	0.077*	0.050
	[0.0458]	[0.0490]	[0.0506]	[0.0429]	[0.0433]	[0.0451]
Female	–0.035*	–0.027	–0.033	–0.071***	–0.056***	–0.058***
	[0.0195]	[0.0215]	[0.0217]	[0.0205]	[0.0210]	[0.0212]
Married	–0.001	0.004	0.002	–0.037*	–0.019	–0.018
	[0.0208]	[0.0219]	[0.0224]	[0.0213]	[0.0214]	[0.0218]
Education: High school and above	–0.053**	0.007	0.006	–0.067***	–0.011	–0.003
	[0.0244]	[0.0255]	[0.0264]	[0.0256]	[0.0264]	[0.0270]
HH wealth/1,000,000	–0.413***	–0.269**	–0.281***	–0.331***	–0.208**	–0.183*
	[0.1068]	[0.1067]	[0.1070]	[0.0987]	[0.0967]	[0.0951]
Poor/fair general health status		0.105***	0.103***		0.136***	0.132***
		[0.0240]	[0.0243]		[0.0238]	[0.0243]
BMI		0.019***	0.020***		0.019***	0.018***
		[0.0018]	[0.0018]		[0.0017]	[0.0017]
Cognition		–0.005**	–0.006**		–0.004*	–0.004*
		[0.0023]	[0.0024]		[0.0023]	[0.0024]
Ever smoker		0.021	0.025		0.010	0.007
		[0.0208]	[0.0209]		[0.0207]	[0.0209]
Current smoker		–0.066**	–0.067**		–0.035	–0.046
		[0.0322]	[0.0326]		[0.0336]	[0.0337]
Vigorous exercise		–0.051**	–0.052**		–0.036*	–0.035
		[0.0211]	[0.0212]		[0.0218]	[0.0219]
Enrolled in part B			0.074**			0.055
			[0.0351]			[0.0375]
Drug coverage (part D or other sources)			0.041			0.086***
			[0.0270]			[0.0300]
Medicaid			0.012			0.097**
			[0.0443]			[0.0424]
Covered by EGHP			0.007			0.031
			[0.0202]			[0.0205]
Observations	2,517	2,172	2,127	2,369	2,221	2,177

Notes: Excluding individuals who have Medicare coverage through an HMO. Average marginal effects after probit estimation. Dependent variable = 1 if individual diabetic according to any of the three measures (self-reports, biomarker, or claims).

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 7.11 Probability of self-reported diabetes—Marginal effects after probit estimation

	2006			2008		
	SES	+health indicators	+insurance status	SES	+health indicators	+insurance status
HbA1c Level	0.211***	0.187***	0.187***	0.185***	0.167***	0.165***
	[0.0082]	[0.0087]	[0.0087]	[0.0074]	[0.0077]	[0.0077]
Age	–0.002**	–0.001	–0.002	–0.003***	–0.002	–0.002
	[0.0009]	[0.0012]	[0.0012]	[0.0009]	[0.0012]	[0.0012]
White	–0.008	0.012	0.009	–0.007	0.018	0.014
	[0.0214]	[0.0221]	[0.0228]	[0.0208]	[0.0205]	[0.0212]
Hispanic	0.010	–0.006	–0.009	0.040	0.002	–0.010
	[0.0319]	[0.0331]	[0.0341]	[0.0303]	[0.0291]	[0.0294]
Female	–0.018	–0.016	–0.019	–0.062***	–0.054***	–0.056***
	[0.0151]	[0.0168]	[0.0170]	[0.0151]	[0.0156]	[0.0158]
Married	–0.002	–0.005	–0.005	–0.021	–0.016	–0.016
	[0.0160]	[0.0171]	[0.0175]	[0.0154]	[0.0157]	[0.0161]
Education: High school and above	–0.029	0.009	0.007	–0.057***	–0.027	–0.031
	[0.0188]	[0.0193]	[0.0200]	[0.0191]	[0.0196]	[0.0204]
HH wealth/1,000,000	–0.083	–0.009	–0.018	–0.192**	–0.120	–0.110
	[0.0755]	[0.0764]	[0.0775]	[0.0807]	[0.0781]	[0.0773]
Poor/fair general health status		0.088***	0.090***		0.064***	0.063***
		[0.0194]	[0.0197]		[0.0178]	[0.0181]
BMI		0.010***	0.011***		0.007***	0.007***
		[0.0014]	[0.0014]		[0.0012]	[0.0012]
Cognition		–0.003	–0.003*		–0.003**	–0.003*
		[0.0018]	[0.0018]		[0.0017]	[0.0018]
Ever smoker		0.013	0.014		0.012	0.011
		[0.0161]	[0.0163]		[0.0151]	[0.0153]
Current smoker		–0.043*	–0.044*		–0.039*	–0.037
		[0.0246]	[0.0250]		[0.0226]	[0.0230]
Vigorous exercise		–0.028*	–0.026		0.010	0.011
		[0.0811]	[0.0167]		[0.0165]	[0.0166]
Enrolled in part B			0.037			–0.003
			[0.0274]			[0.0280]
Drug coverage (part D or other sources)			0.034*			0.054***
			[0.0207]			[0.0211]
Medicaid			0.030			0.045
			[0.0361]			[0.0314]
Covered by EGHP			0.017			0.023
			[0.0159]			[0.0152]
Observations	2,517	2,172	2,127	2,325	2,178	2,135

Notes: Excluding individuals who have Medicare coverage through an HMO. Average marginal effects after probit estimation. Dependent variable = 1 if individual has self-reported diabetes.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

3. Undiagnosed3: HbA1c level > = 6.5 percent, but no self-reported diabetes, no diabetes in claims data, and no glucose/HbA1c screening test two years before the interview.

The first measure uses only information from self-reports and biomarker data. Individuals are coded as having undiagnosed diabetes if their HbA1c level is above 6.5 percent, but they do not report having been diagnosed with diabetes, that is, this definition ignores information from claims. In the second definition we incorporate the information from the claims data: only individuals who have an elevated HbA1c level but neither report diabetes in the HRS nor have diabetes related claims are coded as having undiagnosed diabetes. In the third definition we further require that individuals have not taken a screening test in the two years before the HRS interview.

The last three columns of tables 7.7 and 7.8 show means of the different demographic, socioeconomic, and health-related variables for the three different definitions of undiagnosed diabetes. Compared to individuals with diabetes, fewer among the undiagnosed are Hispanic, a larger share has at least a high school degree, they have higher average income, higher median wealth (although the mean is lower in 2006), and fewer rate their health as fair or poor. These findings are somewhat surprising—as they suggest that, if at all, individuals with higher SES have a higher risk of undiagnosed diabetes. In order to shed additional light on this, the final set of regressions studies these relationships in multivariate analyses.

In the multivariate analyses of predictors of undiagnosed diabetes, the undiagnosed cases are compared to those who have diagnosed diabetes. In the first definition, a diagnosis of diabetes can come from self-reports or biomarkers, in the second and third definitions, individuals who only have diabetes according to the claims data are also coded as diabetic. Results for the first definition are presented in table 7.12, for the second in table 7.13, and for the third in table 7.14. Interestingly, fair or poor self-rated health is the only predictor of undiagnosed diabetes that is robust and significant across all definitions and in almost all specifications. It is conceivable that a diabetes diagnosis leads individuals to rate their health as fair or poor, so the question of causality has to be left open. In addition, in some of the specifications some of the socioeconomic or health-related variables significantly predict undiagnosed diabetes. However, the patterns are not consistent across years. Overall, there is thus no systematic relationship of any of our measures of undiagnosed diabetes with demographic or socioeconomic characteristics. Neither is there an impact of cognition, known risk factors, such as BMI or exercise, or health insurance status. Given the richness of our data and findings in the earlier literature we reviewed above, this is perhaps a bit surprising.

Table 7.12 **Probability of undiagnosed diabetes (undiagnosed1)**

	2006			2008		
	SES	+health indicators	+insurance status	SES	+health indicators	+insurance status
Age	0.003	0.001	0.001	0.005**	0.002	0.002
	[0.0017]	[0.0024]	[0.0024]	[0.0019]	[0.0026]	[0.0026]
White	–0.063	–0.063	–0.071	–0.038	–0.052	–0.056
	[0.0397]	[0.0468]	[0.0485]	[0.0399]	[0.0451]	[0.0466]
Hispanic	–0.052	–0.063	–0.053	–0.020	0.015	0.030
	[0.0414]	[0.0461]	[0.0491]	[0.0498]	[0.0609]	[0.0665]
Female	0.006	–0.004	0.001	0.091***	0.075**	0.078**
	[0.0286]	[0.0336]	[0.0334]	[0.0293]	[0.0321]	[0.0326]
Married	–0.002	0.003	–0.001	0.038	0.028	0.029
	[0.0297]	[0.0333]	[0.0336]	[0.0307]	[0.0328]	[0.0335]
Education: High school and above	0.023	0.004	–0.002	0.047	0.030	0.022
	[0.0300]	[0.0375]	[0.0389]	[0.0319]	[0.0378]	[0.0393]
HH wealth/1,000,000	–0.228	–0.326	–0.295	0.013	–0.037	–0.023
	[0.2262]	[0.2656]	[0.2559]	[0.1711]	[0.1907]	[0.1831]
Poor/fair general health status		–0.088***	–0.092***		–0.075**	–0.075**
		[0.0305]	[0.0307]		[0.0324]	[0.0329]
BMI		–0.001	–0.001		–0.004	–0.005*
		[0.0028]	[0.0028]		[0.0027]	[0.0027]
Cognition		–0.001	0.001		0.002	0.002
		[0.0035]	[0.0036]		[0.0038]	[0.0039]
Ever smoker		–0.043	–0.040		–0.045	–0.041
		[0.0332]	[0.0330]		[0.0332]	[0.0334]
Current smoker		0.065	0.066		0.043	0.041
		[0.0710]	[0.0716]		[0.0637]	[0.0638]
Vigorous exercise		0.013	0.006		–0.030	–0.031
		[0.0351]	[0.0343]		[0.0347]	[0.0350]
Enrolled in part B			–0.011			–0.011
			[0.0634]			[0.0608]
Drug coverage (part D or other sources)			–0.054			–0.005
			[0.0505]			[0.0572]
Medicaid			–0.052			–0.056
			[0.0485]			[0.0464]
Covered by EGHP			–0.025			–0.043
			[0.0304]			[0.0320]
Observations	635	542	532	635	584	575

Notes: Excluding individuals who have Medicare coverage through an HMO. Average marginal effects after probit estimation. Sample includes individuals identified as diabetics in self-reports or biomarker data. Dependent variable = 1 if diabetes in biomarker data, but not according to self-reports.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 7.13 Probability of undiagnosed diabetes (undiagnosed2)

	2006			2008		
	SES	+health indicators	+insurance status	SES	+health indicators	+insurance status
Age	–0.000	–0.001	–0.001	–0.000	–0.001	–0.001
	[0.0012]	[0.0017]	[0.0017]	[0.0013]	[0.0017]	[0.0017]
White	–0.003	–0.001	–0.002	–0.036	–0.053	–0.064*
	[0.0248]	[0.0289]	[0.0298]	[0.0299]	[0.0344]	[0.0368]
Hispanic	–0.033	–0.043	–0.035	–0.051*	–0.042	–0.030
	[0.0281]	[0.0290]	[0.0328]	[0.0262]	[0.0315]	[0.0379]
Female	0.012	0.011	0.013	0.041**	0.025	0.026
	[0.0198]	[0.0225]	[0.0225]	[0.0202]	[0.0221]	[0.0225]
Married	–0.007	–0.006	–0.010	0.036*	0.026	0.022
	[0.0210]	[0.0231]	[0.0235]	[0.0213]	[0.0227]	[0.0233]
Education: High school	0.018	0.002	–0.002	0.017	–0.006	–0.017
and above	[0.0211]	[0.0267]	[0.0279]	[0.0227]	[0.0281]	[0.0302]
HH wealth/1,000,000	–0.180	–0.256	–0.204	0.018	–0.002	0.000
	[0.1851]	[0.2146]	[0.1947]	[0.1000]	[0.1172]	[0.1209]
Poor/fair general health		–0.048**	–0.043**		–0.034	–0.033
status		[0.0204]	[0.0207]		[0.0219]	[0.0225]
BMI		–0.000	–0.000		–0.001	–0.001
		[0.0019]	[0.0020]		[0.0018]	[0.0018]
Cognition		0.001	0.002		0.004	0.003
		[0.0025]	[0.0025]		[0.0026]	[0.0027]
Ever smoker		–0.003	–0.004		–0.046**	–0.046*
		[0.0219]	[0.0220]		[0.0233]	[0.0236]
Current smoker		–0.022	–0.016		0.029	0.033
		[0.0345]	[0.0367]		[0.0459]	[0.0478]
Vigorous exercise		0.009	0.004		–0.004	–0.008
		[0.0239]	[0.0235]		[0.0241]	[0.0244]
Enrolled in part B			–0.008			–0.005
			[0.0471]			[0.0447]
Drug coverage (part D or			0.019			0.017
other sources)			[0.0282]			[0.0356]
Medicaid			–0.043			–0.055**
			[0.0288]			[0.0271]
Covered by EGHP			–0.024			–0.014
			[0.0206]			[0.0219]
Observations	801	698	683	837	776	761

Notes: Excluding individuals who have Medicare coverage through an HMO. Average marginal effects after probit estimation. Sample includes individuals identified as diabetics in self-reports, biomarker, or claims data. Dependent variable = 1 if diabetes in biomarker data, but not according to self-reports and claims.

***Significant at the 10 percent level.

**Significant at the 5 percent level.

*Significant at the 1 percent level.

Table 7.14 **Probability of undiagnosed diabetes (undiagnosed3)**

	2006			2008		
	SES	+health indicators	+insurance status	SES	+health indicators	+insurance status
Age	–0.001	–0.002	–0.002	–0.000	–0.001	–0.001
	[0.0011]	[0.0016]	[0.0016]	[0.0012]	[0.0016]	[0.0016]
White	–0.012	–0.008	0.000	–0.019	–0.028	–0.037
	[0.0237]	[0.0269]	[0.0262]	[0.0269]	[0.0303]	[0.0325]
Hispanic	–0.018	–0.028	–0.028	–0.044*	–0.036	–0.028
	[0.0276]	[0.0283]	[0.0283]	[0.0241]	[0.0288]	[0.0339]
Female	0.009	0.008	0.007	0.034*	0.018	0.018
	[0.0180]	[0.0203]	[0.0204]	[0.0191]	[0.0210]	[0.0214]
Married	–0.007	–0.012	–0.010	0.028	0.017	0.015
	[0.0190]	[0.0212]	[0.0213]	[0.0202]	[0.0214]	[0.0220]
Education: High school and above	0.013	0.000	0.005	0.001	–0.024	–0.032
	[0.0193]	[0.0247]	[0.0245]	[0.0224]	[0.0282]	[0.0300]
HH wealth/1,000,000	–0.086	–0.122	–0.121	0.014	–0.005	–0.005
	[0.1435]	[0.1599]	[0.1566]	[0.0961]	[0.1120]	[0.1118]
Poor/fair general health status		–0.047***	–0.044**		–0.045**	–0.045**
		[0.0183]	[0.0186]		[0.0201]	[0.0206]
BMI		0.000	–0.000		–0.000	–0.001
		[0.0017]	[0.0018]		[0.0017]	[0.0017]
Cognition		0.001	0.001		0.002	0.002
		[0.0023]	[0.0023]		[0.0024]	[0.0025]
Ever smoker		–0.001	0.001		–0.047**	–0.047**
		[0.0198]	[0.0199]		[0.0219]	[0.0222]
Current smoker		–0.008	–0.009		0.004	0.006
		[0.0335]	[0.0334]		[0.0402]	[0.0418]
Vigorous exercise		–0.003	–0.002		–0.002	–0.005
		[0.0209]	[0.0210]		[0.0227]	[0.0230]
Enrolled in part B			–0.014			0.003
			[0.0427]			[0.0404]
Drug coverage (part D or other sources)			0.004			0.003
			[0.0277]			[0.0357]
Medicaid						–0.045*
						[0.0259]
Covered by EGHP			–0.008			–0.020
			[0.0188]			[0.0205]
Observations	801	698	686	837	776	761

Notes: Excluding individuals who have Medicare coverage through an HMO. Average marginal effects after probit estimation. Sample includes individuals identified as diabetics in self-reports, biomarker, or claims data. Dependent variable = 1 if diabetes in biomarker data, but not according to self-reports and claims, and has not taken a diabetes screening two years before the HRS interview.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

7.4 Summary and Outlook

In this chapter we compare three measures of diabetes using HRS data: the commonly used survey measure on diabetes, diabetes according to HbA1c levels collected in the HRS biomarker data, and diabetes in the Medicare insurance claims linked to the HRS data. Self-reported diabetes and diabetes information from biomarker data align for a large part of our sample (85 percent). Using information on self-reported medication from the HRS as well as information from claims data help to shed light on the differences between the self-reports and the biomarker data. Most of the differences can likely be explained by the fact that treatment lowers HbA1c levels in some cases even below the 6.5 percent threshold. When considering the three data sources, roughly 2–3 percent of individuals have diabetes according to HbA1c but do not report diabetes, and do not receive diabetes treatment according to their claims records. Even in the Medicare population there is thus a fraction of individuals who likely have undiagnosed diabetes. Somewhat surprisingly, however, we do not find that the probability of being undiagnosed is related to socioeconomic status.

Importantly, comparing the three measures of diabetes as well as taking into account information on treatment suggests that none of the three measures should be taken as a gold standard. In particular, our results stress that both the presumably more objective biomarker as well as the claims data suffer from error just as the self-reports. While the biomarker data can be influenced by treatment and thus may not identify cases as diabetic because their diabetes is well managed, the claims data may potentially falsely classify individuals as diabetics (e.g., Sakshaug, Weir, and Nicholas 2014). In addition, individuals who have diabetes but are not treated for it will also be misclassified based on the claims data.

We envision that future research will move beyond the descriptive analysis of the data we presented in this chapter. A statistical model could start from a framework (e.g., Wansbeek and Meijer 2000) in which true disease prevalence is unobserved, with survey self-reports, biomarkers, and administrative claims data being three indicators that all potentially suffer from measurement error. Such a model could be used to construct a more reliable measure of prevalence, which in turn could be employed as a predictor in substantive analysis, for example, mortality prediction or studies of health care use.

Another issue that future research might address is that typically not all respondents of a survey provide consent to biomarker measurement or administrative record linkage. Also, while biomarker data and administrative linkages potentially improve measurement of disease prevalence in community surveys, they involve costs as well. Both these issues could be addressed jointly in a statistical decision framework motivated by a total survey error cost perspective (Groves 1989). Specifically, this approach could address the questions of whether collecting biomarkers or administrative linkage are worth their costs, which of them is the more cost effective, and whether including both of them is the best option.[9] The required cost and benefit calculations are, however, more straightforward for a biomarker such as HbA1c as its only purpose is the measurement of diabetes prevalence, while linked insurance claims data can serve many purposes so that their benefits are harder to quantify. To end on a positive note, one of our results was that adding claims information to combined self-reports and biomarkers reduces undiagnosed diabetes cases from 3.26 percent to 2.4 percent in 2006 and from 4.05 percent to 3.1 percent in 2008, that is, by between one-quarter and one-third. Thus, including all three measures in a major study such as the HRS improves measurement of disease prevalence substantially.

References

American Diabetes Association. 2010. "Standards of Medical Care in Diabetes—2010." *Journal of Clinical and Applied Research in Education* 33 (Supp. 1): 11–61.

———. 2015. "Standards of Medical Care in Diabetes—2015." *The Journal of Clinical and Applied Research in Education* 38 (Supp. 1): 1–94.

Baker, Michael, Mark Stabile, and Catherine Deri. 2004. "What Do Self-Reported, Objective, Measures of Health Measure?" *Journal of Human Resources* 39:1067–93.

Barcellos, Silvia, Dana Goldman, and James P. Smith. 2012. "Undiagnosed Disease, Especially Diabetes, Casts Doubt on Some of Reported Health 'Advantage' of Recent Mexican Immigrants." *Health Affairs* 31:2727–37.

Bennett, C. M., M. Guo, and S. C. Dharmage. 2007. "HbA1c as a Screening Tool for Detection of Type 2 Diabetes: A Systematic Review." *Diabetic Medicine* 24:333–43.

Bonora, Enzo, and Jaako Tuomilehto. 2011. "The Pros and Cons of Diagnosing Diabetes with A1c." *Diabetes Care* 34 (Supp. 2): 184–90.

Chatterji, Pinka, Heesoo Joo, and Kajal Lahiri. 2012. "Beware of Being Unaware: Racial/Ethnic Disparities in Chronic Illness in the USA." *Health Economics* 21:1040–60.

Goldman, Noreen, I-Fen Lin, Maxine Weinstein, and Yu-Hsuan Lin. 2003. "Evaluating the Quality of Self-Reports of Hypertension and Diabetes." *Journal of Clinical Epidemiology* 56:148–54.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Johnston, David, Carol Propper, and Michael Shields. 2009. "Comparing Subjective and Objective Measures of Health: Evidence from Hypertension for the Income/Health Gradient." *Journal of Health Economics* 28:540–52.

Manski, Charles F., and Francesca Molinari. 2008. "Skip Sequencing: A Decision Problem in Questionnaire Design." *Annals of Applied Statistics* 2:264–85.

[9] We are aware of only one paper that formally studies costs and benefits of decisions in survey design, although in a different context (Manski and Molinari 2008).

Reynolds, Timothy, Stuart Smellie, and Patrick Twomey. 2006. "Glycated Haemoglobin (HbA1c) Monitoring." *British Medical Journal* 333 (7568): 586–88.

Rohlfing, C. L., R. R. Little, H. H. Wiedmeyer, J. D. England, R. Madsen, M. I. Harris, K. M. Flegal, M. S. Eberhardt, and D. E. Goldstein. 2000. "Use of GhB (HbA1c) in Screening for Undiagnosed Diabetes in the US Population." *Diabetes Care* 23:187–91.

Sakshaug, Joseph W., David R. Weir, and Lauren H. Nicholas. 2014. "Identifying Diabetics in Medicare Claims and Survey Data: Implications for Health Services Research." *BMC Health Services Research* 14 (150): 1–6.

Smith, James P. 2007. "Nature and Causes of Trends in Male Diabetes Prevalence, Undiagnosed Diabetes, and the Socioeconomic Status Health Gradient." *Proceedings of the National Academy of Sciences of the United States of America* 204:13225–31.

Wansbeek, Tom, and Erik Meijer. 2000. *Measurement Error and Latent Variables in Econometrics*. Amsterdam: Elsevier.

Wolinsky, Frederik D., Michael P. Jones, Fred Ullrich, Yiyue Lou, and George L. Wehby. 2014. "The Concordance of Survey Reports and Medicare Claims in a Nationally Representative Longitudinal Cohort of Older Adults." *Medical Care* 52 (5): 462–68.

Yasaitis, Laura, Lisa Berkman, and Amitabh Chandra. 2015. "Comparison of Self-Reported and Medicare Claims-Identified Acute Myocardial Infarction." *Circulation* 131:1477–85.

Comment James P. Smith

In a thought-provoking chapter, Heiss et al. raise several important questions about the appropriate way to measure diabetes prevalence in household surveys. While diabetes is the disease at issue in the chapter, the same questions would arise with many other disease outcomes. Three common measures of diabetes prevalence are used and compared in their analysis—self-reports of ever being diagnosed by a doctor, the common HbA1c diabetes biomarker being above the standard American threshold of 6.5 percent, and a diabetes diagnosis mentioned in Medicare claims data. The question the authors ask is whether the three measures are "consistent" and which one is "correct."

Figure 7C.1, derived from the chapter, illustrates the central finding of the chapter by showing diabetes prevalence rates for a sample of HRS respondents who had their diabetes measured in all three ways in 2006 and in 2008. Rates of diabetes prevalence are clearly quite different using the three

James P. Smith holds the Distinguished Chair in Labor Market and Demographic Studies at the RAND Corporation.

These comments were written with the support of grants from the National Institute of Aging and were delivered at a NBER conference on the Economics of Aging, Carefree, Arizona, May 2015. For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see http://www.nber.org/chapters/c13638.ack.